GLOBAL ENSEMBLE FORECAST SYSTEM PRECIPITATION FORECASTS AND THE IMPLICATIONS OF STATISTICAL DOWNSCALING OVER THE WESTERN UNITED STATES

by

Wyndam Robert Lewis

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Atmospheric Sciences

The University of Utah

December 2016

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of **Wyndam Robert Lewis**

has been approved by the following supervisory committee members:

William James Steenburgh	, Chair	**08-19-2016** Date Approved
Lawrence Bennet Dunn	, Member	**08-19-2016** Date Approved
Courtenay Strong	, Member	**08-19-2016** Date Approved

and by **Kevin D. Perry** , Chair/Dean of

the Department/College/School of **Atmospheric Sciences**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Operational medium-range ensemble modeling systems produce quantitative precipitation forecasts (QPFs) that provide guidance for weather forecasters, yet these systems lack sufficient resolution to adequately resolve orographic influences on precipitation. In this study, we verify cool-season (Oct-Mar) Global Ensemble Forecast System (GEFS) QPFs using daily (24-h) western United States (U.S.) Snow Telemetry (SNOTEL) observations, which tend to be located at upper elevations where orographic enhancement of precipitation is more pronounced. Results indicate widespread dry biases, which reflect the infrequent production of larger 24-h precipitation events ($\geq$ 22.9 mm in Pacific Ranges and $\geq$ 10.2 mm in the Interior Ranges) relative to observations. Performance metrics, such as equitable threat score (ETS), hit rate, and false alarm ratio, generally worsen from the coast toward the interior. The ensemble spread captures only ~30% of upper-quartile events at Day 5, exhibits poor reliability, and is about as skillful over the interior compared to forecasts using climatological probabilities.

In an effort to improve QPFs without exacerbating computing demands, we explore statistical downscaling based on high-resolution climatological precipitation analyses from the Parameter-elevation Relationships on Independent Slopes Model (PRISM), an approach frequently used by operational forecasters. Such downscaling improves model biases, ETSs, and hit rates. However, 50% of downscaled QPFs for

upper-quartile events are false alarms at Day 1, and probabilistic QPFs still do not capture ~40% of such events at Day 5. These results should help forecasters and hydrologists understand the capabilities and limitations of GEFS forecasts and statistical downscaling over the western U.S. and other regions of complex terrain.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Jim Steenburgh, for giving me this opportunity, and providing guidance in a way that has enhanced my skills as a scientist. I would also like to thank my committee members, Larry Dunn and Court Strong, for their valuable input.

I am appreciative of the resources and opportunities made available by the Department of Atmospheric Sciences at the University of Utah, and all those who have contributed to making this experience as beneficial and constructive as possible.

I would like to thank the National Oceanic and Atmospheric Administration (NOAA) Physical Science Division (PSD), the Natural Resources Conservation Service, the PRISM Climate Group at Oregon State University, and the NOAA/National Centers for Environmental Predication, especially Yuejian Zhu and Yan Luo, for providing data used in this study. I would also like to thank the University of Utah Center for High Performance Computing for computer-support services. Trevor Alcott, Tim Barker, Keith Brill, Randy Graham, Glen Merrill, Jon Rutz, and Darren Van Cleave provided comments and suggests that aided the research and writing of this thesis.

This study is supported by the NOAA/National Weather Service CSTAR Program Grant NA13NWS4680003.

CHAPTER 1

INTRODUCTION

1.1 Quantitative Precipitation Forecasts

Accurate quantitative precipitation forecasts (QPFs) in mountainous regions are particularly challenging for meteorologists using current operational ensemble prediction systems, which lack sufficient resolution to adequately resolve critical convective and orographic processes that strongly influence the distribution and intensity of precipitation over complex terrain (Junker et al. 1992; Kunz and Kottmeier 2006; Smith et al. 2010; Haren et al. 2015). Over the western United States (U.S.), for example, meteorologists must infer how local terrain features will modulate rainfall and snowfall, as well as precipitation impacts on air and ground transportation, water resource and flood management, outdoor recreation, and avalanche safety (Stewart et al. 1995; Cohen 1996; Neiman et al. 2001; Ralph et al. 2006; U.S. Dept. of the Interior 2012; Black and Mote 2015; Schirmer and Jamieson 2015; Parker and Abatzoglou 2016). Knowledge of model biases, capabilities, and limitations has the potential to improve forecasts, but is limited by a scarcity of studies evaluating operational ensemble model performance in areas of complex terrain (Shirmer and Jamieson 2015).

The majority of western U.S. precipitation occurs during the cool season, defined here as October through March, with a significant portion falling as snow at higher

elevations (Serreze et al. 1999). Large precipitation events with high snow levels, many associated with atmospheric rivers (ARs), yield hydrological extremes that produce flooding, property and infrastructure damage, and loss of life (Neiman et al. 2001; Ralph et al. 2006; U.S. Dept. of the Interior 2012; Rutz and Steenburgh 2014). For example, orographic enhancement during AR conditions produced all seven major floods on California's Russian River from October 1997 to February 2006 (Ralph et al. 2006). Many mountain areas and highways in the western U.S. are also susceptible to avalanche hazards, with snowfall and rainfall increasing the likelihood of natural and human-triggered avalanches (Tremper 2008; Hatchett and Kaplan 2016). State Route 210 in Utah, for example, crosses 50 avalanche paths and is hit by an average of 33 avalanches per year (Steenburgh 2014). Winter-precipitation-related motor vehicle- and aviation-related accidents result in roughly 900 fatalities on average each year across the U.S., with some of the highest standardized mortality rates occurring in the west (Black and Mote 2015). Nationally, such fatalities amount to more than double the combined fatalities from lightning, tornados, hurricanes, heat, and cold (Black and Mote 2015).

QPFs are typically more skillful in the cool season when large-scale dynamic forcing dominates precipitation generation, as opposed to the localized convection and weaker dynamic forcing found during the warm season (Junker et al. 1992; Mullen and Buizza 2000; Baxter et al. 2014). Nevertheless, QPF skill is often lower in mountainous regions, due at least in part to poorly resolved terrain features (Junker et al. 1992; Yuan et al. 2005; Ikeda et al. 2010) and, over the western U.S. interior, low spatial coherence of precipitation events (Serreze et al. 2001; Parker and Abatzoglou 2016). Complex terrain also impedes QPFs by contributing to cyclone displacement errors that skew QPF

positioning and timing (Charles and Colle 2009), and similar effects are likely to affect the track and moisture transport of large-scale weather systems, including atmospheric rivers (e.g., Rutz et al. 2014, 2015).

This study focuses on the Global Ensemble Forecast System (GEFS), an operational ensemble modeling system run by the U.S. National Weather Service (NWS) that is widely used by forecasters in the western U.S. With an effective horizontal grid spacing of ~33 km, the GEFS is unable to resolve key topographical features and subsequent effects on precipitation (NOAA 2015). While we are unaware of any peer-reviewed analyses examining the performance of the current GEFS, which became operational in December 2015, Hamill (2011) showed that an earlier version of the GEFS produced probabilistic QPFs (PQPFs) with insufficient spread, lower reliability, and lower Brier skill scores compared to the European, Canadian, and United Kingdom ensemble modeling systems. Baxter et al. (2014) also evaluated an earlier version of the GEFS, showing that GEFS QPFs have little useful skill over the southeast U.S. by forecast day 5.5 (108-132 h), and that GEFS PQPFs demonstrate little to no skill compared to climatological reference probabilities by forecast day 6.5 (132-156 h).

There are several approaches targeted at improving QPFs from coarse-resolution ensembles, such as calibration (Eckel and Walters 1998), dynamic downscaling (e.g., Stensrud et al. 1999; Marsigli et al. 2001), bias correction and statistical disaggregation (Wood et al. 2002), Bayesian Model Averaging (Raftery et al. 2005), analog sorting (Bontron and Obled 2005), and reforecast analogs (Hamill and Whitaker 2006). In this study, we use statistical downscaling, which is computationally inexpensive, widely employed in climate and hydrological applications (e.g., Wilby et al. 1998; Wood et al.

2004; Gutmann et al. 2012), and used by many NWS Forecast Offices and River Forecast Centers in the western U.S. Such downscaling often uses high-resolution (~800-m grid spacing) precipitation analyses produced by the PRISM Climate Group at Oregon State University (Daly et al. 1994, 2008) to rescale lower resolution model guidance and provide increased spatial detail. Described in greater depth in Section 2.2, the approach implicitly assumes climatological precipitation distributions and that the small-scale precipitation variability is directly related to the large-scale precipitation pattern. This yields climatologically plausible precipitation distributions, but may be problematic during storms that are strongly influenced by unresolved mesoscale processes (e.g., mesoscale precipitation bands, non-orographic convection, etc.) or feature precipitation-altitude relationships that deviate from climatology (e.g., Steenburgh 2003, 2004), particularly in regions where orographic enhancement is sensitive to flow direction.

1.2 Scope of This Study

The purpose of this study is to provide a comprehensive overview of GEFS QPF and PQPF performance over the western U.S., including an evaluation of statistical downscaling using high-resolution PRISM climatology. Specifically, we examine the performance of the current operational version of the GEFS relative to the CPC Unified Daily Precipitation Analysis (hereafter CPC analysis) and upper-elevation Snow Telemetry (SNOTEL) observations. These datasets and the methods used for evaluation are described in Chapter 2. Results are then presented in Chapter 3, with conclusions and a discussion of the significance of our findings provided in Chapter 4.

CHAPTER 2

DATA AND METHODS

2.1 Global Ensemble Forecast System

We verify reforecasts (i.e., retrospective forecasts) and forecasts produced by the current (as of 2 December 2015) operational version of the GEFS, which is based on version 12.1.0 of the National Centers for Environmental Prediction (NCEP) Global Spectral Model [the forecast component of the Global Forecast System (GFS)], configured with 64 vertical levels and a horizontal resolution of TL574 (~33 km) for the first 192 h and TL382 (~55 km) from 192–384 h (NOAA 2015). Ensemble members consist of a control and 20 perturbations generated with an ensemble Kalman filter scheme. Reforecasts for the 2013/14 and 2014/15 cool seasons were obtained from the NCEP NOMADS server (Rutledge et al. 2006), whereas reforecasts (1 October to 1 December 2015) and forecasts for the remainder of the 2015/16 cool season were provided by the NCEP Environmental Modeling Center. Although GEFS forecasts are currently provided four times a day on a 0.5º lat-lon grid, we use 0000 UTC initialized runs on a 1.0º lat-lon grid since this is the only initialization time and output grid spacing available for the 2013/14 and 2014/15 reforecast periods. Given that the GEFS is typically available a few hours after the nominal initialization time, we define Day 1 as the 12-36 h forecast and perform validation through Day 7 (156-180 h), which

concentrates on the higher resolution portion of GEFS forecasts.

2.2 Downscaling Methodology

The climatology-based statistical downscaling method used here is similar to that employed at NWS Forecast Offices, River Forecast Centers, and the Weather Prediction Center (WPC) and uses monthly, climatological (1981-2010) high-resolution (30-arcsec, ~800-m grid spacing) precipitation analyses produced by the PRISM Climate Group at Oregon State University [analysis technique described by Daly et al. (1994)]. First, we generate a daily precipitation climatology for the forecast day of interest by interpolating the monthly PRISM precipitation analyses to daily values and calculating a centered 15-day average (Fig. 2.1a). We then smooth the daily values to a spatial scale consistent with the GEFS 1.0º lat-lon grid (Fig. 2.1b). In operational practice, a variety of techniques are used for this smoothing, including Gaussian filtering at WPC (Keith Brill, WPC, personal communication) and grid aggregation by averaging at many NWSFOs (Tim Barker, NWSFO Boise, personal communication). We selected a Gaussian filter with a full width of 1.0º. Results will, however, vary some depending on smoothing or aggregation technique and the scale over which it is applied. Next, we divide the original PRISM precipitation analysis by the Gaussian-smoothed analysis to obtain an analysis of *downscaling ratio* across the western U.S. (Fig. 2.1c).

Bilinearly interpolating the GEFS QPF (Fig. 2.1d) to the PRISM grid (Fig. 2.1e) and multiplying by the downscaling ratio yields the downscaled QPF (Fig. 2.1f). The downscaling ratio is typically less than one in valleys and basins, leading to a downscaled QPF that is lower than the GEFS QPF. Conversely, the downscaling ratio is typically

greater than one in mountains and upland regions, leading to a downscaled QPF that is larger than the GEFS QPF. For point verification in this study, GEFS QPFs and daily downscaling ratios are bilinearly interpolated directly to observation locations and multiplied to obtain downscaled QPFs.

2.3 Precipitation Analyses and Observations

For gridded validation, we use the NOAA/Climate Prediction Center (CPC) Unified Daily Precipitation Analysis (hereafter the CPC analysis) on a 0.25º lat-lon grid (Higgins et al. 2000; Xie et al. 2007; Chen et al. 2008) and bilinearly interpolate GEFS QPFs to the CPC analysis grid for comparison. Although higher resolution precipitation analyses are available [e.g., the Climatology-Calibrated Precipitation Analysis (Hou et al. 2014)], the lower resolution CPC analysis is sufficient for identifying broad regional biases in GEFS forecasts.

Gauge-based verification in upper-elevation regions uses accumulated (since 0000 PST 1 October) precipitation observations from the Snowpack Telemetry (SNOTEL) network maintained by the National Resources Conservation Service (NRCS). The automated SNOTEL stations measure precipitation collected by a large-storage weighing gauge in imperial units at 0.1 inch precision. SNOTEL stations are typically strategically placed in sheltered areas with regionally high snow accumulations and include an Alter wind shield to reduce undercatch (Yang et al. 1998; Serreze et al. 1999; Fassnacht 2004). Comparable gauges have shown an undercatch of ~10-15% for wind speeds of about 1-2 m s^{-1} (Yang et al. 1998; Fassnacht 2004), which is a typical wind speed found in forest clearings that house SNOTEL stations (Ikeda et al. 2010). Additional factors influencing

SNOTEL precipitation data include transmission errors, instrument malfunction (e.g., leaks), temperature-based fluctuations (affecting readings by the pressure transducer), and snow adhesion to gauge walls (delaying precipitation measurement). See Serreze et al. (1999) and Avanzi et al. (2014) for summaries of the capabilities and limitations of SNOTEL measurements.

Instrument limitations warrant our implementation of basic quality control to reduce the use of erroneous data. We begin quality control with hourly cumulative precipitation observations downloaded from the NRCS, identifying negative values (typically -99.9 and -0.1 in). If these values are surrounded by equal non-negative values, we replace the negative values with the surrounding non-negative value, otherwise they are flagged as erroneous. We then discretely sample the 1200 UTC observations and identify spikes of more than (less than) 0.5 inches above (below) the maximum (minimum) of the surrounding 20 days. If these spikes are surrounded by equal values, we replace them with the equal value, otherwise they are flagged as erroneous. After these adjustments, we calculate the daily (1200–1200 UTC) precipitation for all periods when data are available for the current and prior day, setting all negative values to zero and flagging all values in excess of 5.0 inches as erroneous. The latter removes many false jumps in the data, along with a small sample of actual extreme events. The number of these extreme events is, however, too small for statistically significant results. Daily precipitation values valid at 1200 UTC 1 October require accumulated precipitation data from the previous water year, and are not included.

After these checks, we remove stations that contain erroneous data on 20% of the days during the three cool seasons. Then, for each remaining station, we calculate the

ratio of cumulative daily precipitation during the three cool seasons relative to that obtained from the gauge's accumulated measurement at the end of each cool season. We then remove stations at which this ratio is 1 (1.5) standard deviation above (below) the median ratio for all stations. The more relaxed criterion for lower ratios reflects the removal of events > 5 inches from the daily precipitation. Daily precipitation data are then converted from inches to millimeters. These requirements result in data from 603 of 781 stations being used for the validation.

2.4 Verification Methods

No single statistical measure can adequately diagnose strengths and weaknesses of a numerical weather prediction model (Schaefer 1990). We use a series of measures based on a 2x2 contingency table commonly used for precipitation validation (Table 2.1), to provide a broad assessment of the capabilities of the GEFS and downscaled GEFS. These measures are described in Joliffe and Stephenson (2003) and include:

$$Hit\ rate = \frac{a}{a+b}, \tag{2.1}$$

$$False\ alarm\ ratio = \frac{c}{a+d}, \tag{2.2}$$

$$Bias\ score = \frac{a+b}{a+c}, \tag{2.3}$$

and

$$Equitable\ threat\ score\ (ETS) = \frac{a-a_r}{a-a_r+c+b}, \tag{2.4}$$

where

$$a_r = \frac{(a+c)(a+b)}{n}. \tag{2.5}$$

The hit rate is equal to the fraction of correct forecasts (hits) to observed events. The false alarm ratio expresses the fraction of forecasts that do not verify as events. The bias score represents the fraction of forecasts issued to events observed. The equitable threat score (ETS) is a common precipitation verification tool for two-category (dichotomous) events, providing a single value between 1 (perfect forecast) and 0 (equivalent to a random forecast) (Jolliffe and Stephenson 2003; Hamill and Juras 2006).

We acknowledge the existence of varying SNOTEL station precipitation climatologies in subsequent ETS calculations, which tends to result in over-estimated ETSs, and therefore adopt modifications as outlined in Hamill and Juras (2006). Rather than calculate ETS across the western U.S. based on one contingency table, ETS is calculated as a weighted average generated from 10 subgroups with similar climatological reference probabilities (defined as the fraction of daily-precipitation observations that exceed a defined threshold). Specifically, the modified ETS calculation follows

$$\overline{\mathrm{ETS}} = \sum_{s=1}^{10} u(s)\,\mathrm{ETS}(s), \tag{2.6}$$

where $u(s)$ denotes the fraction of SNOTEL stations in each subgroup, s, which are approximately of equal size.

Probabilistic verification utilizes reliability diagrams (illustrating the relation of forecast probabilities to their observed frequencies), Brier skill scores (BSS; a measure of probabilistic forecast skill relative to climatological reference probabilities), rank diagrams (indicate where observations fall within the ensemble spread), and additional forecast attributes to help gauge the overall value of the GEFS (Jolliffe and Stephenson 2003). To account for variations in climatological reference probabilities across stations

(Wilks 2006, Chapter 7; Hamill et al. 2007), reliability diagrams include a histogram inset that displays the frequency of occurrence of forecast probabilities and the SNOTEL climatological reference probabilities in 10% bins. We also use resampling to generate 5% and 95% consistency bars, which indicate the variability among observed frequencies due to limited counting statistics (Jolliffe and Stephenson 2003; Brocker and Smith 2007). The approach is similar to the bootstrapping methods in Hamill et al. (2007) and follows a technique known as consistency resampling (Brocker and Smith 2007). We resample 1000 times, using $2N$ samples rather than N^2 samples since N is O(10^5). See Brocker and Smith (2007) for details.

Similar to ETS, we calculate the BSS as the weighted average generated from 10 subgroups with similar climatological reference probabilities. Specifically, the BSS is calculated following

$$\mathrm{BSS} = \sum_{s=1}^{10} u(\mathrm{s}) \left[1 - \frac{\overline{BS}^{\mathrm{f}}(\mathrm{s})}{\overline{BS}^{\mathrm{c}}(\mathrm{s})} \right], \tag{2.7}$$

where $u(s)$ denotes the fraction of SNOTEL stations in each subgroup, s, $\overline{BS}^{\mathrm{f}}(\mathrm{s})$ represents the average Brier score (mean squared error of a probabilistic forecast; see Wilks 2006, Chapter 7) for forecasts in subgroup s, and $\overline{BS}^{\mathrm{c}}(\mathrm{s})$ represents climatological reference probabilities in subgroup s.

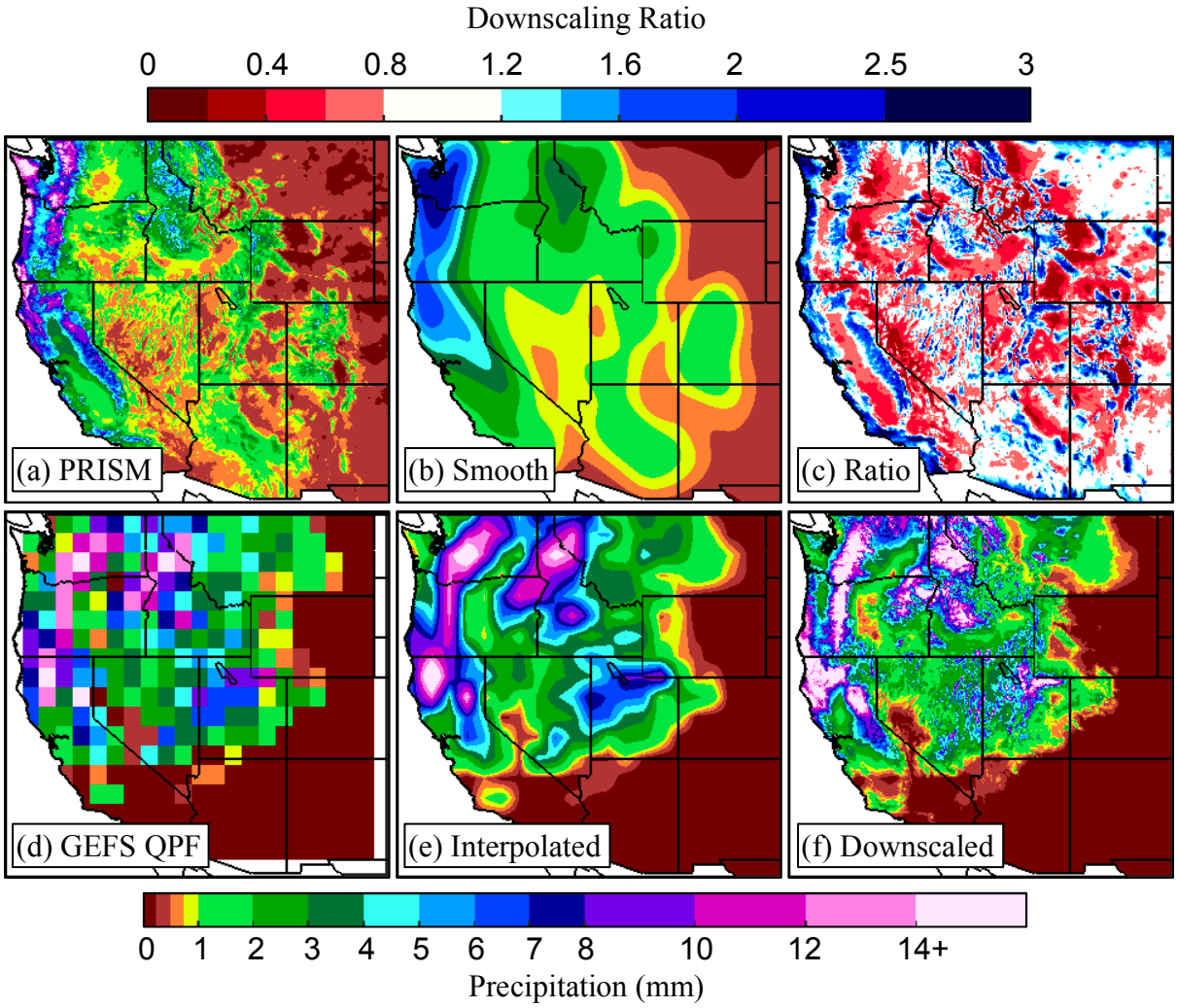

Fig. 2.1. Statistical downscaling example for 24 Jan 2016. (a) Daily PRISM-derived climatological precipitation. (b) same as (a) except smoothed with the Gaussian filter. (c) Downscaling ratio obtained by dividing (a) by (b). (d) GEFS Day 1 control forecast on 1° lat-lon grid. (e) GEFS Day 1 control forecast bilinearly interpolated to PRISM grid. (f) Downscaled forecast after multiplying (c) and (e).

Table. 2.1. Contingency table used for forecast validation.

		Observed	
Forecast		Yes	No
	Yes	(a) Hit	(b) False alarm
	No	(c) Miss	(d) Correct rejection

CHAPTER 3

RESULTS

3.1 GEFS Climatology

We begin by comparing mean-daily precipitation in the CPC analysis to that produced by the GEFS Day 1 control forecast to describe the climate of the three cool-season study period and identify regional-scale climatological biases in the GEFS. Biases and other forecast characteristics exhibited by the GEFS control are similar to those of the other individual GEFS members. During the three cool-season study period, CPC-analyzed precipitation was heaviest in the coastal ranges of the Pacific Northwest and northern California and the Cascade Mountains of Washington and Oregon (Fig. 3.1a). Over the interior, precipitation was heaviest over regions with higher terrain including northern and central Idaho, northwest Montana, north-central Utah, western Colorado, and the Mogollon Rim. The interior northwest was wetter than the interior southwest, which reflects both climatology and persistent drought conditions over the latter. The GEFS Day 1 control captures the broad regional characteristics of the CPC precipitation distribution (Fig. 3.1c); however, the ratio of GEFS control to CPC precipitation (i.e., the bias ratio) reveals that the GEFS control is too dry over and upstream of topographic barriers and too wet in downstream valleys and basins (Fig. 3.1e). When comparing the GEFS control at SNOTEL stations (Fig. 3.1b) to SNOTEL observations (Fig. 3.1d), a dry

bias is evident at most stations (Fig. 3.1f), which are located preferentially in the mountains. At 23% (61%) of the SNOTEL stations, the bias ratio is smaller than 0.5 (0.75), indicating a substantial dry bias. By Day 5, the GEFS control bias ratio relative to both the CPC analysis and SNOTEL stations has shifted to slightly lower values, revealing a tendency for the GEFS control (as well as other individual GEFS members) to become drier with increasing forecast lead time (cf. Figs. 3.1e, 3.2a and 3.1f, 3.2b).

A comparison of the frequency of daily (24-h) precipitation (2.54 mm bins) produced by the GEFS with that at CPC analysis grid points (Fig. 3.3a) and SNOTEL stations (Fig. 3.3b) identifies biases in event frequency as a function of event size. Compared to the CPC analysis, which spans the low and high elevations of the western U.S., the GEFS Day 1 control produces too many events ≥ 20.3 mm and too few events ≥ 22.9 mm (Fig. 3.3a). The largest frequency bias (forecast/observed) is associated with 5.1 mm events, above which the frequency bias exhibits a near monotonic decline with increasing event size (Fig. 3.3a). For all western U.S. SNOTEL stations, which are located primarily at upper elevations, events ≥ 7.6 mm are predicted at a frequency consistent with observations, while events ≥ 10.2 mm are associated with an underprediction of event frequency that worsens with increasing event size (Fig. 3.3b). Consideration of undercatch, as might be expected with SNOTEL gauges (Serreze et al. 1999), would further amplify the underprediction. Similar results are found for Day 5 (bias ratio not shown for clarity). Averaging across all members has little impact at Day 1 when the ensemble spread is small, but averaging at longer lead-times results in an increased number of smaller events and a decrease in the number of larger events, exacerbating these frequency biases (bias ratios for ensemble means also not shown for

clarity).

Distinct regional differences in frequency bias are revealed when grouping SNOTEL stations based on geography, climate, and model performance. We examined several regional groupings, but ultimately present results from two highly differentiated regions: Pacific Ranges and Interior Ranges (Fig. 3.4, stations from intermediate stations not presented for brevity). In the Pacific Ranges, consisting of stations in the Cascade Mountains, Sierra Nevada, and coastal ranges of the Pacific Northwest, the GEFS control produces similar-to-observed event frequencies (i.e., $0.8 \leq$ frequency bias ≤ 1.2) for event sizes ≤ 22.9 mm (Fig. 3.5a). For the Interior Ranges, consisting of inland stations of the Pacific Northwest, Utah, and the Rocky Mountains of Wyoming, Colorado, and New Mexico, similar-to-observed event frequencies are confined to event sizes ≤ 7.6 mm (Fig. 3.5b). Above these thresholds, frequency biases in both regions asymptote toward zero with increasing event size, but are consistently lower in the Interior Ranges, reflective of a larger underprediction bias (cf. Figs. 3.5a,b). We hypothesize that the greater underprediction of event frequency in the Interior Ranges partly reflects the fine scale nature of the topography and inherently low spatial coherence of precipitation systems over the western interior (Serreze et al. 2001).

Bivariate histograms comparing observed and forecast precipitation provide an additional perspective on the GEFS control performance (Fig. 3.6). Skewness in the distribution of more frequent forecast-observation pairs relative to the 1-to-1 line confirms that at all but the smallest thresholds, observed events are more likely than not to be underforecast at Day 1 in the Pacific Ranges (Fig. 3.6a), with the underforecasting worsening over the Interior Ranges (Fig. 3.6b). Collectively, observed events ≥ 15.2 mm

(5.1 mm) in the Pacific (Interior) Ranges at Day 1 are at least twice as likely to be underforecast than overforecast, whereas events ≥ 25.4 mm (12.7 mm) are at least five times as likely to be underforecast. Compared to Day 1, Day 5 forecasts exhibit greater scatter with frequency isolines oriented more normal to the 1-to-1 line, especially in the Interior Ranges, which is consistent with declining skill with increasing forecast projection (Figs. 3.6c,d).

3.2 Downscaled GEFS Climatology

Next, we evaluate the mean-daily precipitation from the downscaled GEFS control relative to SNOTEL observations (Fig. 3.7, see Figs. 3.1b and 3.2b for SNOTEL mean-daily precipitation). At Day 1 and Day 5, downscaling addresses the widespread underprediction evident over mountains in the GEFS control, yielding wetter precipitation climatologies at 93% of SNOTEL stations (cf. Figs. 3.7a, 3.1d and 3.7b, 3.2b). Most of the 7% of stations that become drier with downscaling are on the lee side of the Cascades where GEFS spillover is excessive, and in most instances the drying yields improved bias ratios (i.e., closer to 1). For all western U.S. stations, downscaling increases the median bias ratio at Day 1 from 0.67 to 1.09 and at Day 5 from 0.62 to 1.00. At Day 1, 9%, 53% and 38% of the stations have dry (<0.8), near-neutral (0.8-1.2), and wet (>1.2) bias ratios, respectively, with a greater fraction of stations over the Interior Ranges exhibiting wet bias ratios (Fig 3.7c). Consistent with the GEFS control becoming drier with increasing forecast projection, the downscaled GEFS control bias ratios generally shift slightly to lower values by Day 5 (Fig. 3.7d).

The downscaled GEFS Day 1 control also demonstrates improvements over the

GEFS for event frequency biases at SNOTEL stations. For all western U.S. stations, downscaling produces a relatively consistent 20% overprediction of event frequency for event sizes ≥ 10.2 mm (not shown), which is quite good if one assumes some gauge undercatch. However, there are important regional contrasts in event frequency biases. In the Pacific Ranges, the downscaled GEFS control produces similar-to-observed event frequencies at all event sizes (Fig. 3.8a), whereas there is a clear overprediction of the frequency of events ≥ 10.2 mm in the Interior Ranges that generally worsens with event size.

Bivariate histograms further illustrate that events at Day 1 are less likely to be underforecast by the downscaled GEFS Day 1 control than the undownscaled GEFS control, with the distribution centered closer to the 1-to-1 line over both the Pacific and Interior Ranges, especially for larger events sizes (cf. Figs. 3.6a, 3.9a and 3.6c, 3.9c). However, there is also greater scatter. Median values indicate that at larger event thresholds, an observed event is more likely than not to be underforecast, but, when an event is predicted, it is more likely than not to be an overforecast, especially over the Interior ranges. Like the undownscaled GEFS, downscaled forecasts exhibit little skill by Day 5 (Figs. 3.9b,d).

3.3 Deterministic Validation

Further validation of model performance focuses on upper-quartile precipitation events at CPC grid points and SNOTEL stations. Here, upper quartile is defined as the 75^{th} percentile of observed precipitation events ≥ 2.54 mm (the lowest observable amount at SNOTEL stations) at each grid point or station. Aside from performance measures

inevitably degrading as thresholds are increased, the spatial characteristics of results are generally consistent for other percentile thresholds (e.g., top decile) or absolute precipitation amounts (e.g., 10 mm).

When evaluated using the CPC analysis, GEFS Day 1 control ETSs are generally highest along the Pacific coast and decrease toward the interior with considerable spatial variability (Fig. 3.10a). Compared to SNOTEL observations, GEFS Day 1 control ETSs also exhibit a tendency to decline from the coastal Pacific toward the interior with considerable spatial variability (Figs. 3.10b). ETSs are also generally lower at sites in the interior southwest compared to the interior northwest. These trends are broadly consistent with prior studies that show a positive correlation between ETS and bias (Mason 1989; Hamill 1999), suggesting that the ETS should decline as the dry bias worsens toward the interior. Although this can complicate comparisons of competing models (Hamill 1999), the contrast in ETS between the Pacific and Interior Ranges is statistically significant at a 95% level, suggesting a measurable decline in model skill between the two regions. Downscaling of the GEFS Day 1 control yields ETS improvements at 79% of SNOTEL stations in the western U.S., with the greatest improvements generally over the interior, especially over Utah and Arizona (cf. Figs. 3.10b,c). Although ETSs do increase with downscaling over Montana and Colorado, scores remain relatively low.

Spatial patterns in ETS change minimally with increasing forecast projection, so we instead examine cumulative statistics for upper-quartile events at all SNOTEL stations (Fig. 3.11). Not surprisingly, GEFS control ETSs decline with increasing forecast lead-time, dropping from 0.24 at Day 1 to 0.1, which is sometimes used as a threshold of useful skill (e.g., Baxter et al. 2014), by Day 6 (Fig. 3.11). Downscaling increases ETSs

for all forecast days (Fig. 3.11a), with useful skill extended to Day 7. Based on ETSs, the skill of the downscaled GEFS control at Day 4 is approximately equivalent to the GEFS control at Day 1. ETSs for the GEFS mean (i.e., average of the control plus 20 ensemble members) are slightly worse than the control, whereas the difference between the downscaled control and downscaled GEFS mean is negligible. We suspect that the dry bias of the undownscaled GEFS results in lower ETSs for the GEFS mean compared to the control, especially at longer lead times when the ensemble spread is large. In contrast, the modest wet bias of the downscaled GEFS (e.g., Fig. 3.8) enables comparable ETSs for the GEFS mean and GEFS control at longer lead times.

The underprediction of larger events by the GEFS control is evident in the bias score, with values < 0.6 at all lead times (Fig. 3.11b). Thus, for all SNOTEL stations, the GEFS control produces about half as many upper-quartile events as observed. Downscaling substantially increases the occurrence of larger QPFs, yielding a bias score of 1.2 from Day 1–5 (i.e., 5 forecasted events for every 4 observed), with the bias score declining to just over 1 by Days 6 and 7 (Fig. 3.11b). Bias scores for the GEFS mean and downscaled GEFS mean are slightly lower than the GEFS control and downscaled GEFS control at Day 1, respectively, but decline more rapidly with increasing lead time as the ensemble spread grows and averaging reduces the number of upper-quartile events forecasted (Fig. 3.11b).

The downscaled GEFS control exhibits a much higher hit rate than the control, although hit rates do decrease with increasing lead-time as expected (Fig. 3.11c). At Day 1, the GEFS control upper-quartile hit rate is ~0.3, with downscaling increasing this to ~0.6. However, the downscaled GEFS control also produces more false alarms than the

GEFS control, with a false-alarm ratio at Day 1 of ~0.5 that increases with forecast lead-time (Fig. 3.11d). The GEFS mean produces hit rates and false alarm ratios analogous to the control at short lead times (Figs. 3.11c,d). At longer lead times, the mean produces fewer false alarms but also fewer hits than the control, as averaging an increasing ensemble spread produces fewer upper-quartile events.

Broken down by region, all four of these metrics show a decline in performance from the Pacific Ranges to the Interior Ranges. In the Pacific Ranges, ETSs and hit rates are higher (Figs. 3.12a,c), bias scores are closer to 1 (Fig. 3.12b), and false alarm ratios are lower (Fig. 3.12d) at all lead times. Based on ETS, a Day 5 forecast over the Pacific Ranges is as skillful as a Day 1 forecast over the Interior Ranges (Fig. 3.12a). ETSs for the GEFS control in the Pacific Ranges are even higher than those for the downscaled GEFS control over the Interior Ranges at all forecast lead times, illustrating that even with downscaling, forecast performance is worse over the Interior Ranges than in the Pacific Ranges. The GEFS Day 1 control hit rate for upper-quartile events is 0.44 (0.27) over the Pacific (Interior) Ranges, with a false alarm ratio of 0.27 (0.41) (Figs. 3.12c,d). Downscaling improves Day 1 hit rates to 0.67 (0.59), but worsens false alarm ratios to 0.37 (0.53).

3.4 Probabilistic Validation

Probabilistic validation similarly concentrates on upper-quartile events. We begin by evaluating reliability diagrams, which compare forecast probabilities to their observed frequencies, with close correspondence indicating a reliable ensemble forecast system (Jolliffe and Stephenson 2004). In the Pacific Ranges, reliability diagrams for Day 1

PQPFs exhibit a slope much less than 1, indicating that the GEFS is strongly overconfident (i.e., underdispersive) for short-range forecasting (Fig. 3.13a). Events occur more frequently than predicted when the GEFS produces a low-probability forecast and less frequently than predicted when producing a medium- to high-probability forecast. Similar but somewhat lesser reliability occurs in the Interior Ranges (Fig 3.13b). Reliability over the Pacific Ranges improves through Day 5 when GEFS PQPFs are generally reliable for low-probability forecasts (i.e., < 50%), but still exhibit some overconfidence for high-probability forecasts (cf. Figs. 3.13a,c). Improvement over the Interior Ranges by Day 5 is smaller, and medium- to high-probability forecasts remain strongly overconfident (cf. Figs. 3.13b,d).

Ideally, a probabilistic system exhibits both reliability and sharpness (the relative magnitude of the ensemble spread), with an unreliable yet sharp system being undesirable (Jolliffe and Stephenson 2003). However, in addition to overconfidence, Day 1 GEFS PQPFs are relatively sharp and frequently produce extreme low (0%) and high forecast (100%) probabilities in both the Pacific and Interior Ranges (Figs. 3.13a,b). Sharpness decreases by forecast Day 5 across the western U.S. as extreme low and high forecast probabilities are issued less frequently (Figs. 3.13c,d).

The Brier skill score (BSS) indicates how a probabilistic system performs relative to the climatological reference probability obtained from the sample climatology (Jolliffe and Stephenson 2003). A perfect BSS is 1.0, a BSS of 0.0 indicates no skill over climatology, and a negative BSS indicates skill lower than climatology. BSSs are positive in the Pacific Ranges at Day 1 and Day 5, indicating some skill relative to climatology, although the skill by Day 5 is minimal (Figs. 3.14a,c). BSSs over the Interior Ranges are

only slightly positive on both Day 1 and Day 5, indicating that GEFS PQPFs are about as skillful as climatological probabilities (Figs. 3.14b,d).

Rank histograms illustrate the likelihood of an observation occurring at each location of the ensemble spread when sorted from low to high values (Hamill 2001). Typically, the desired result is that observations are likely to occur between any two ensemble members. While GEFS PQPFs in the Pacific Ranges generally produce a larger ensemble spread and capture 9% (14%) more upper-quartile events at Day 1 (Day 5) than in the Interior Ranges, we present rank histograms for all SNOTEL stations since the underlying theme of results are generally similar.

Consistent with the aforementioned problems predicting larger events, the Day 1 ensemble spread captures only 18% of upper-quartile events, with precipitation amounts during ~80% of those events exceeding the wettest ensemble member (Fig. 3.14a). Upper-quartile events with less precipitation than predicted by the driest ensemble member are relatively rare (3%). Relatively large ensemble spreads are infrequent at Day 1 (Fig. 3.14a), which reflects the sharp and underdispersive nature of the GEFS for short-range forecasting. Larger ensemble spread sizes occur more frequently at longer lead times, such as Day 5 (Fig. 3.14b), allowing the spread to capture 29% of events. However, ~70% of events remain underpredicted by all ensemble members.

Downscaled GEFS Day 1 and Day 5 PQPFs share similar reliability diagram properties compared to the undownscaled GEFS (Fig. 3.13). While downscaling improves the observed occurrence of lower forecast probabilities, higher forecast probabilities are less reliable (Figs. 3.13). Downscaling inherently yields PQPFs that are less sharp due to the enhancement of GEFS QPFs at high-elevation SNOTEL stations

(Fig. 3.13). Downscaling does not improve BSSs over the Interior Ranges at Day 1, and yields relatively small improvements in the Pacific Ranges (Fig. 3.13).

Downscaling reduces sharpness and improves the percent of upper-quartile events captured by the Day 1 ensemble spread from 18% to 40% (cf. Fig. 3.14a, 14c). About 15% of events are overpredicted by all downscaled ensemble members at Day 1, while ~45% are underpredicted (Figs. 3.14c). The downscaled ensemble spread is expectedly further enhanced at Day 5 such that 61% of upper-quartile events are captured, while ~40% remain underpredicted (Figs. 3.14d).

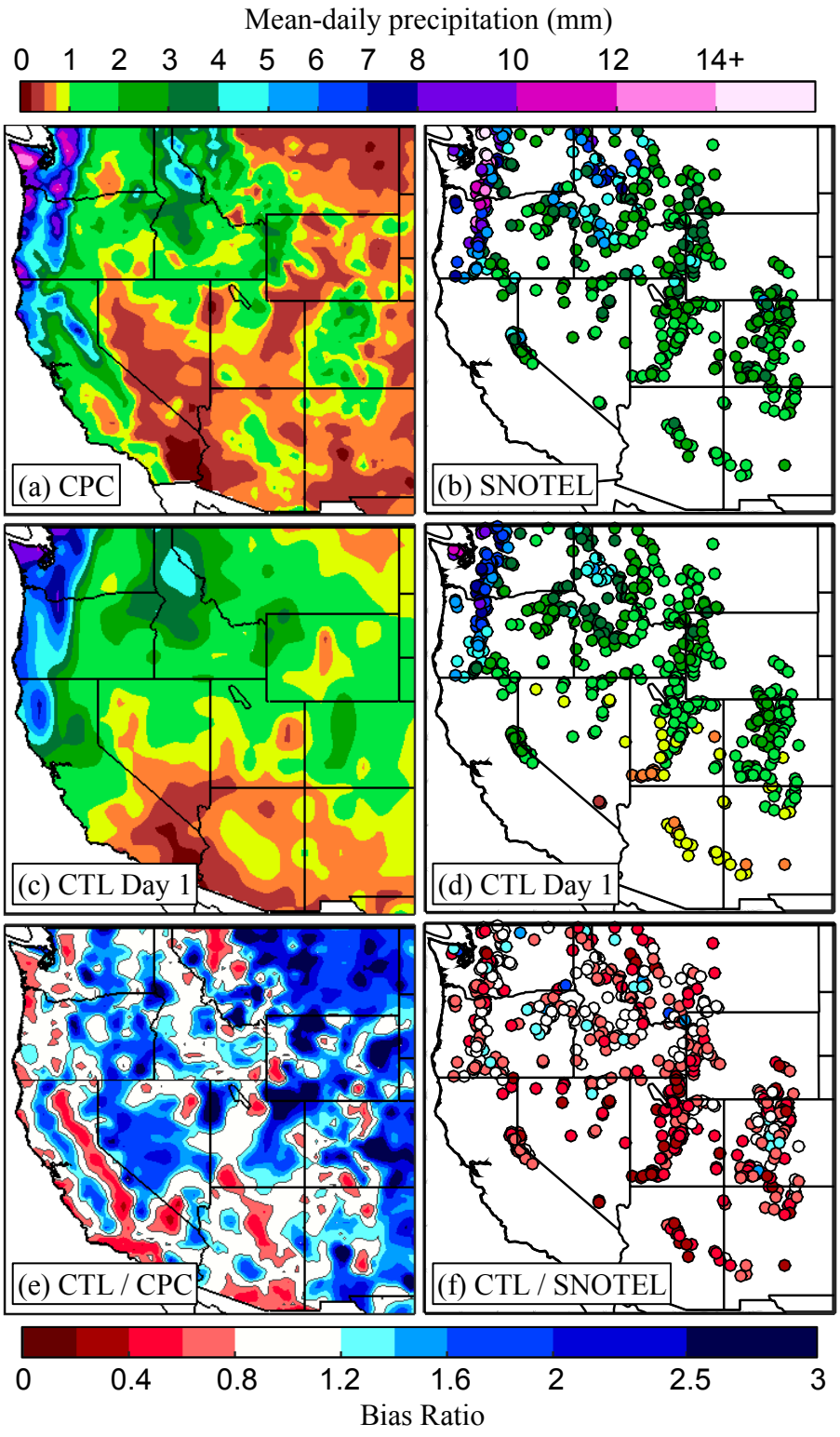

Fig. 3.1. Mean-daily precipitation (mm, upper scale) from the (a) CPC analysis, (b) SNOTEL observations, (c) GEFS Day 1 (12–36 h) control forecast (CTL), and (d) GEFS Day 1 CTL at SNOTEL stations. (e) GEFS Day 1 CTL bias ratio (lower scale) relative to the CPC analysis. (f) Same as (e) except relative to SNOTEL observations.

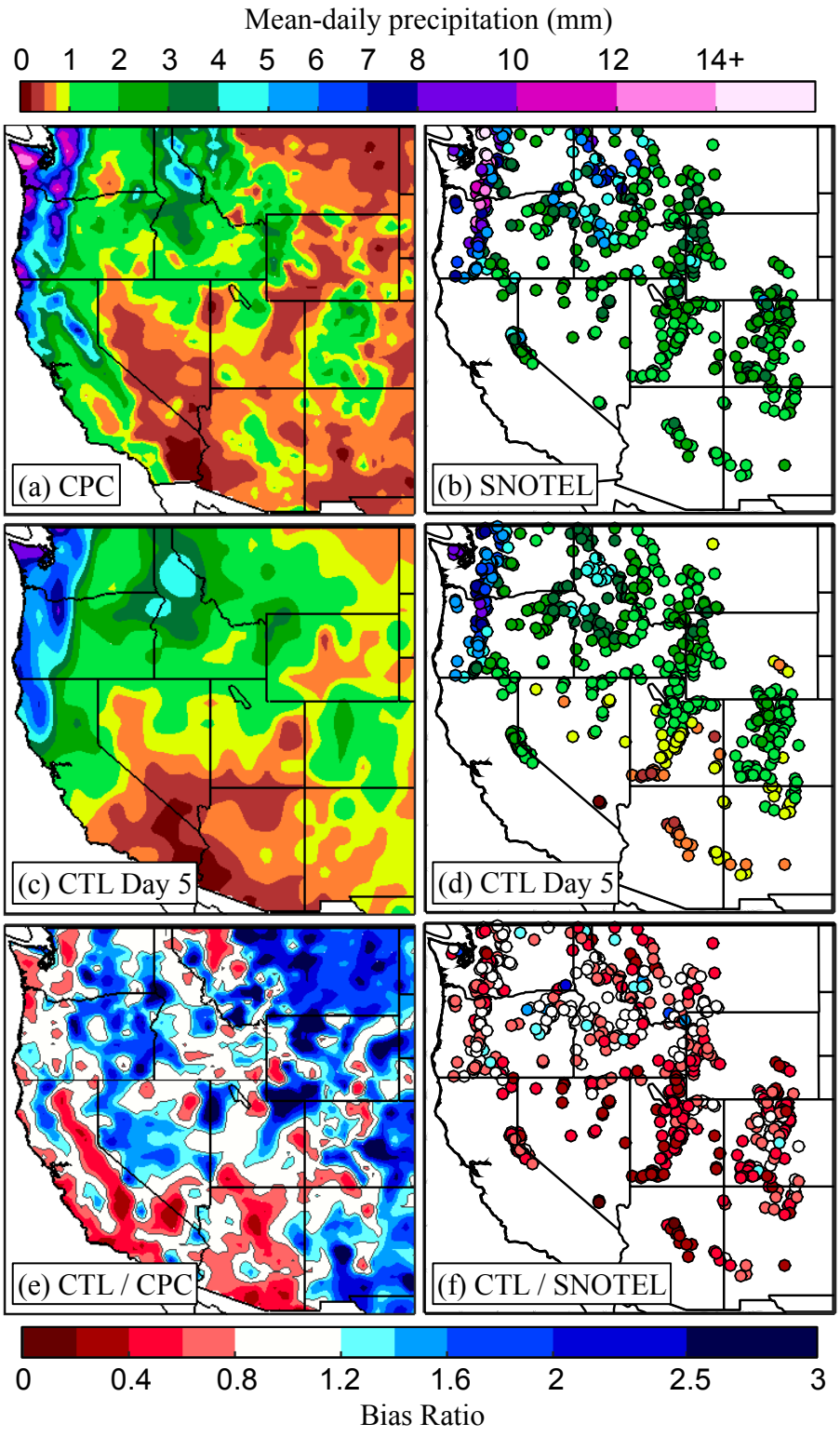

Fig. 3.2. Same as Fig. 3.1 except for Day 5 (108-132 h).

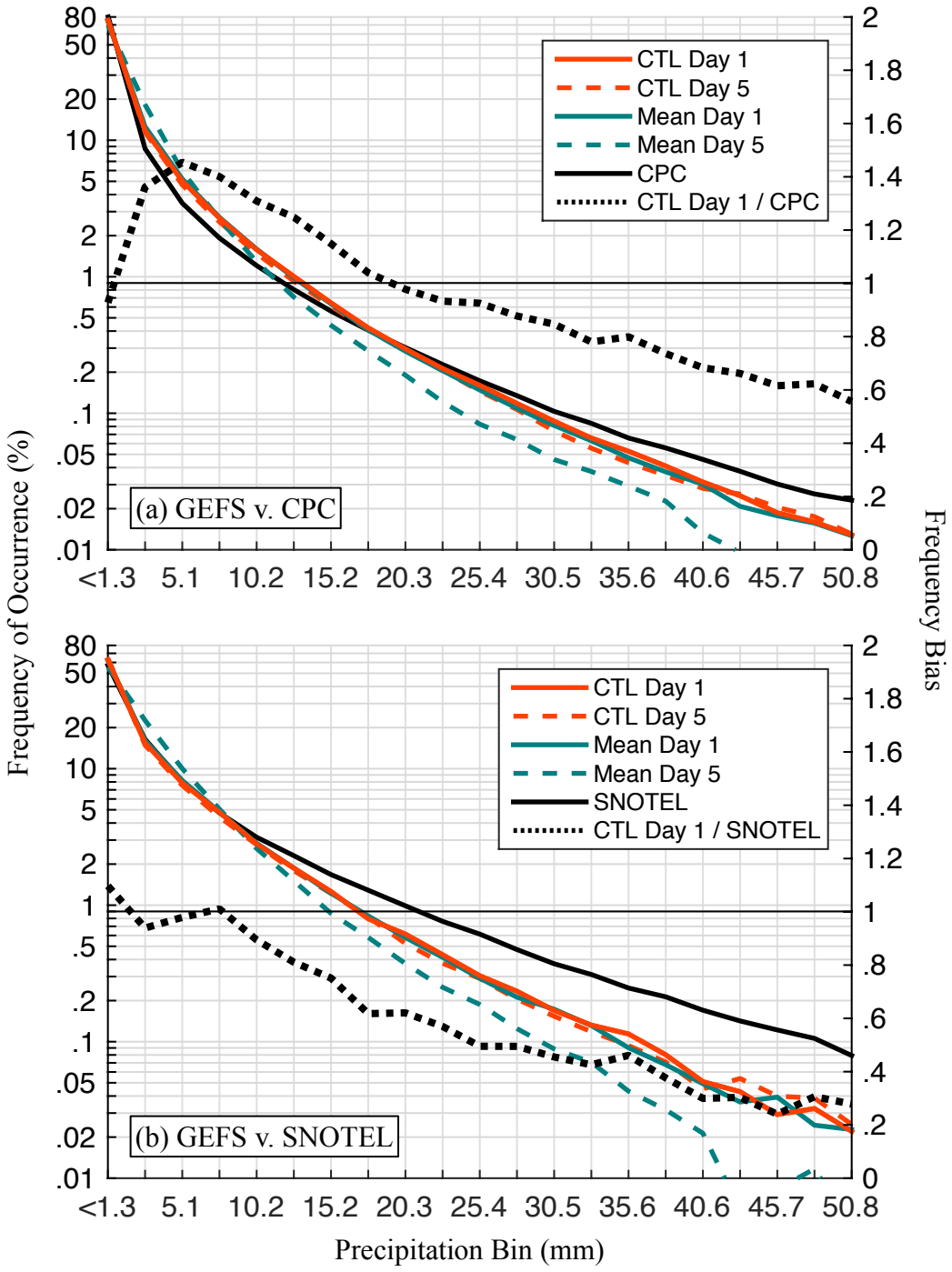

Fig. 3.3. Precipitation event frequency of the (a) GEFS Day 1 (12–36 h) control forecast (CTL Day 1), GEFS Day 5 (108-132 h) CTL (CTL Day 5), GEFS Day 1 ensemble mean forecast (Mean Day 1), GEFS Day 5 Mean (Mean Day 5), and the CPC analysis (CPC) for all CPC analysis grid points in the western U.S. Bias ratio of the GEFS Day 1 CTL to CPC analysis (CTL Day 1 / CPC). (b) Same as (a) except precipitation event frequency at SNOTEL stations (SNOTEL) and bias ratio of the GEFS Day 1 CTL to SNOTEL observations (CTL Day 1 / SNOTEL).

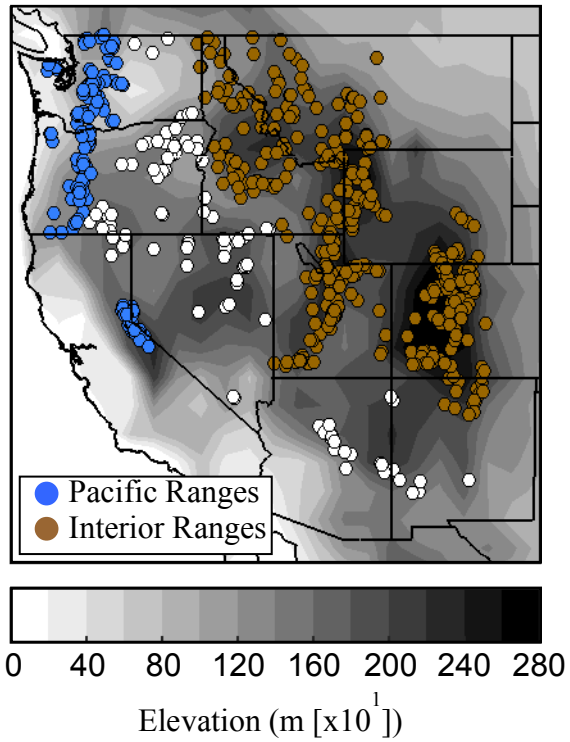

Fig. 3.4. Regional classification of SNOTEL stations with 1ºx1º GEFS topography (shaded following scale at bottom).

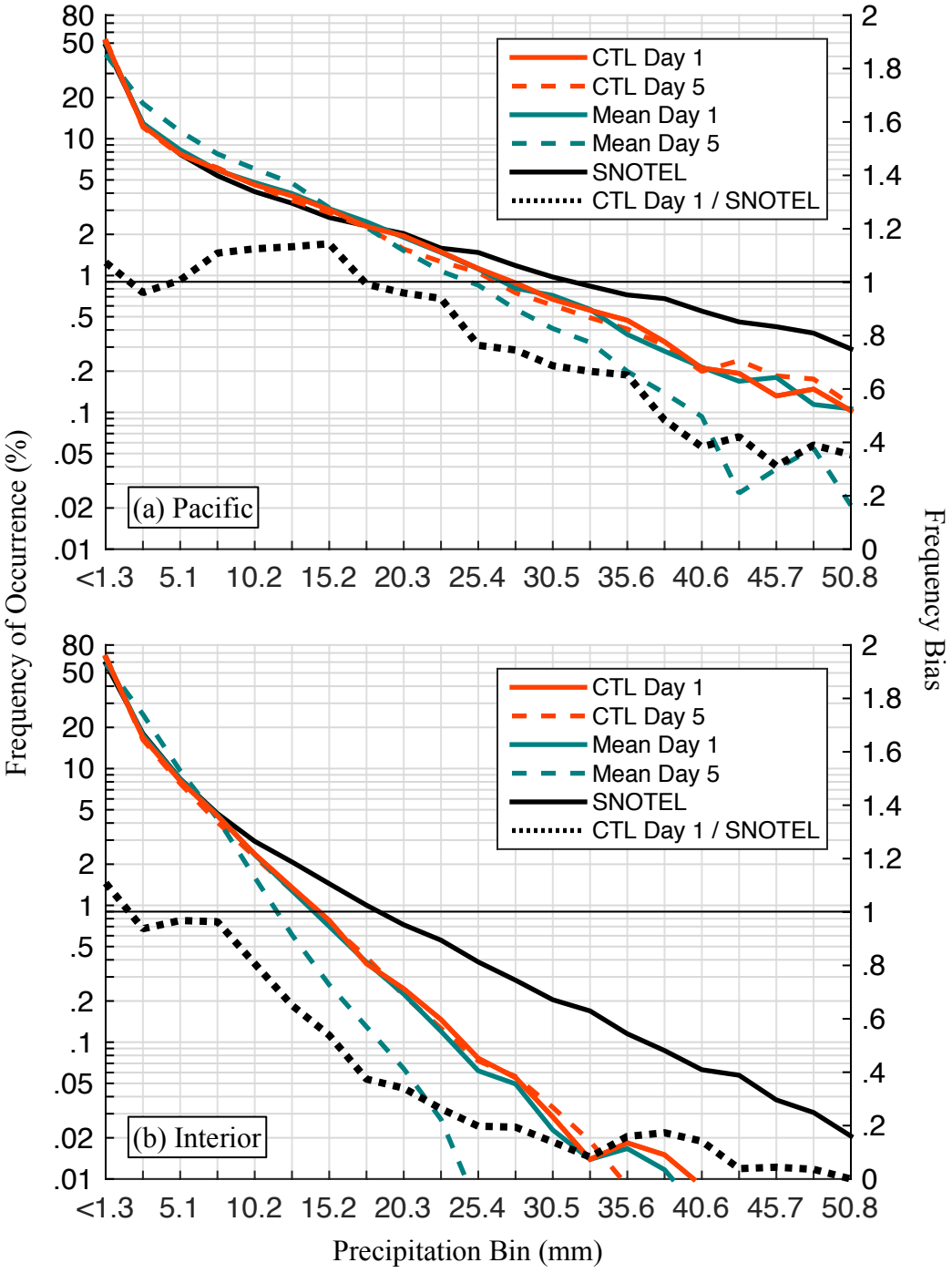

Fig. 3.5. Same as Fig. 3.3b except for (a) Pacific Ranges and (b) Interior Ranges SNOTEL stations.

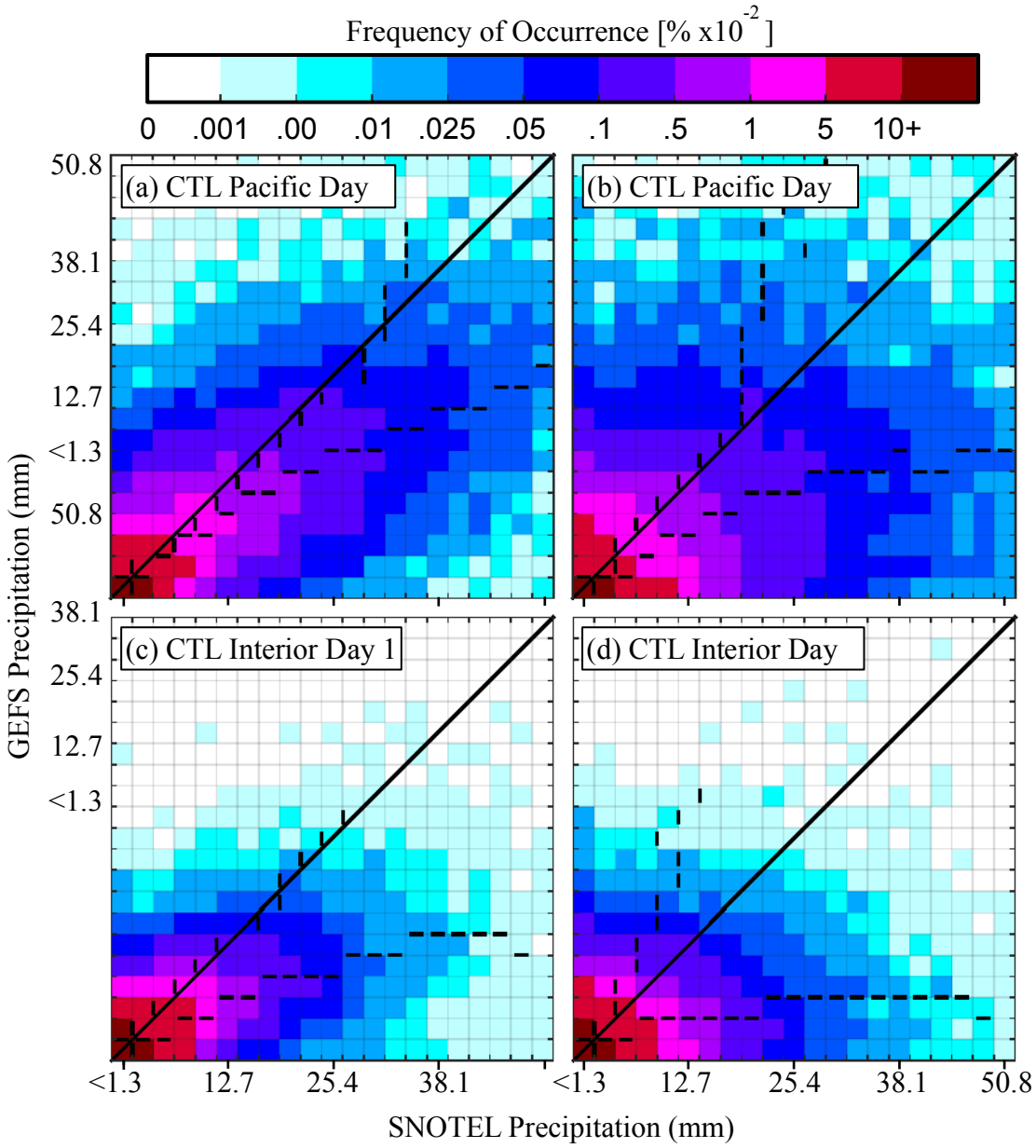

Fig. 3.6. Bivariate histograms of (a) GEFS Day 1 (12-36 h) control (CTL) and observed precipitation at Pacific Ranges SNOTEL stations. (b) Same as (a) except GEFS Day 5 (108-132 h) CTL. (c), (d) Same as (a) and (b) except for Interior Ranges SNOTEL stations. Horizontal (vertical) bars represent the median observed (forecast) value for in each bin.

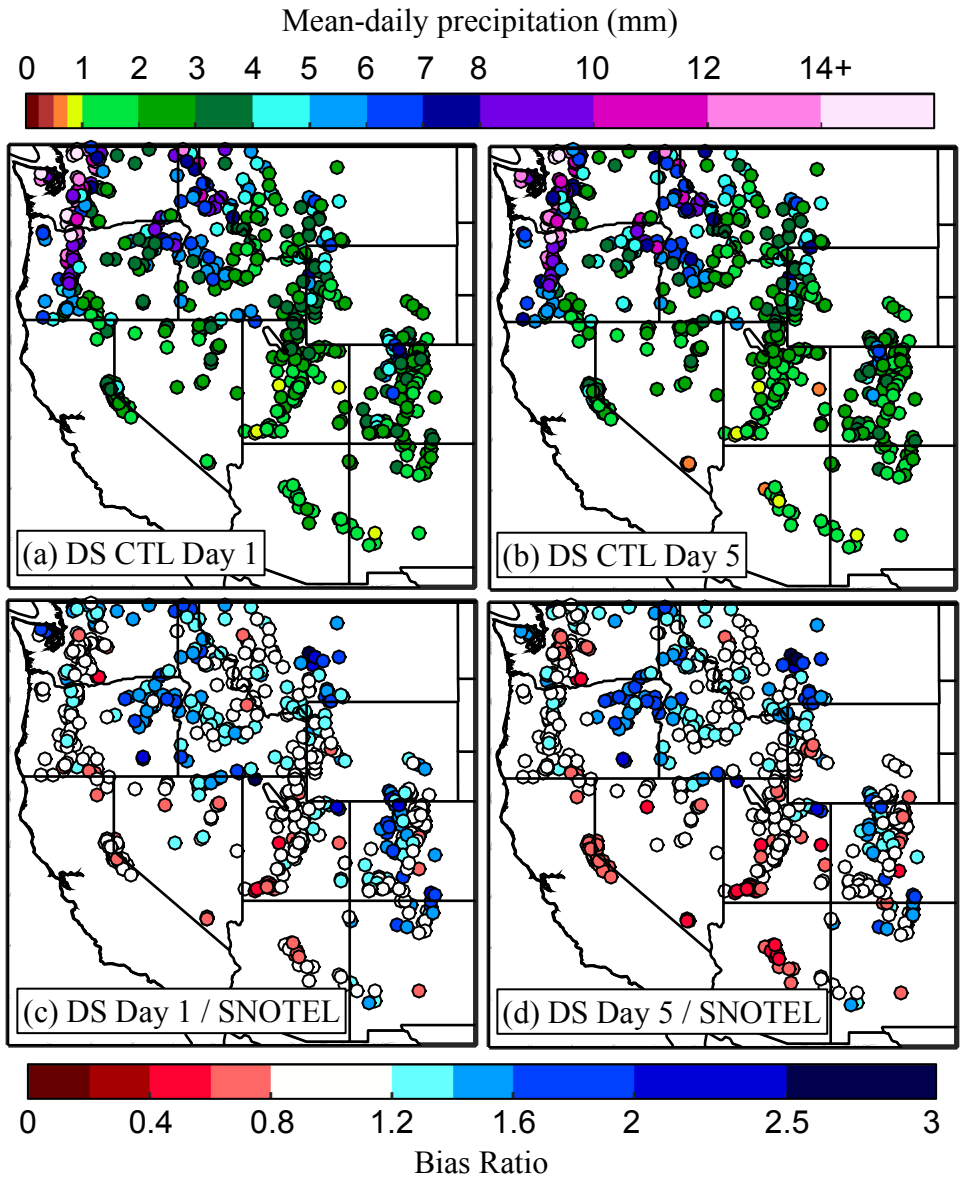

Fig. 3.7. Mean-daily precipitation (mm, upper scale) from the (a) GEFS Day 1 (12–36 h) downscaled control forecast (DS CTL) at SNOTEL stations. (b) Same as (a) except for Day 5 (108-132 h). (c) GEFS Day 1 DS CTL bias ratio (lower scale) relative to SNOTEL observations. (d) Same as (c) except for Day 5.

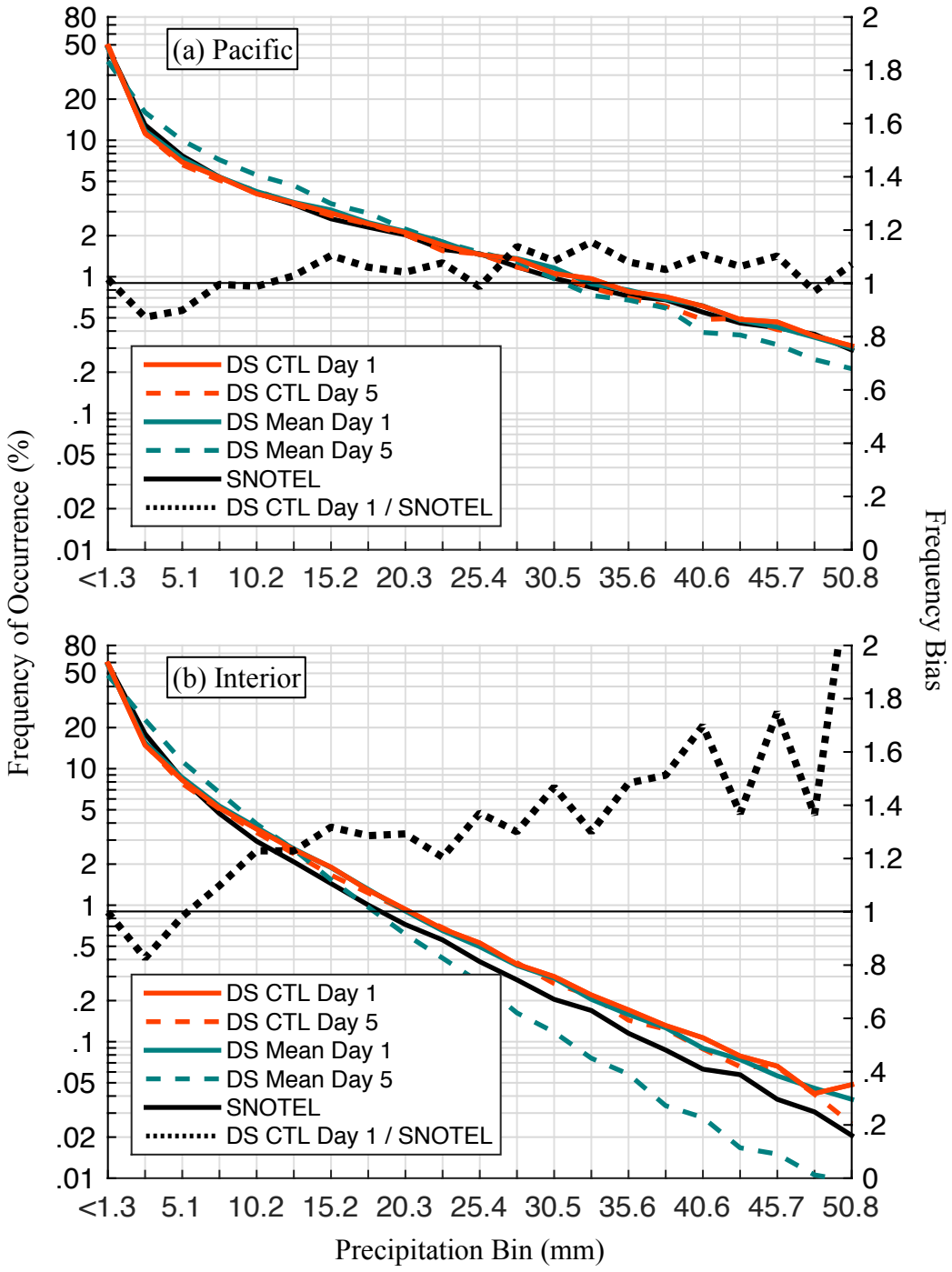

Fig. 3.8. Same as Fig. 3.5 except for the downscaled GEFS.

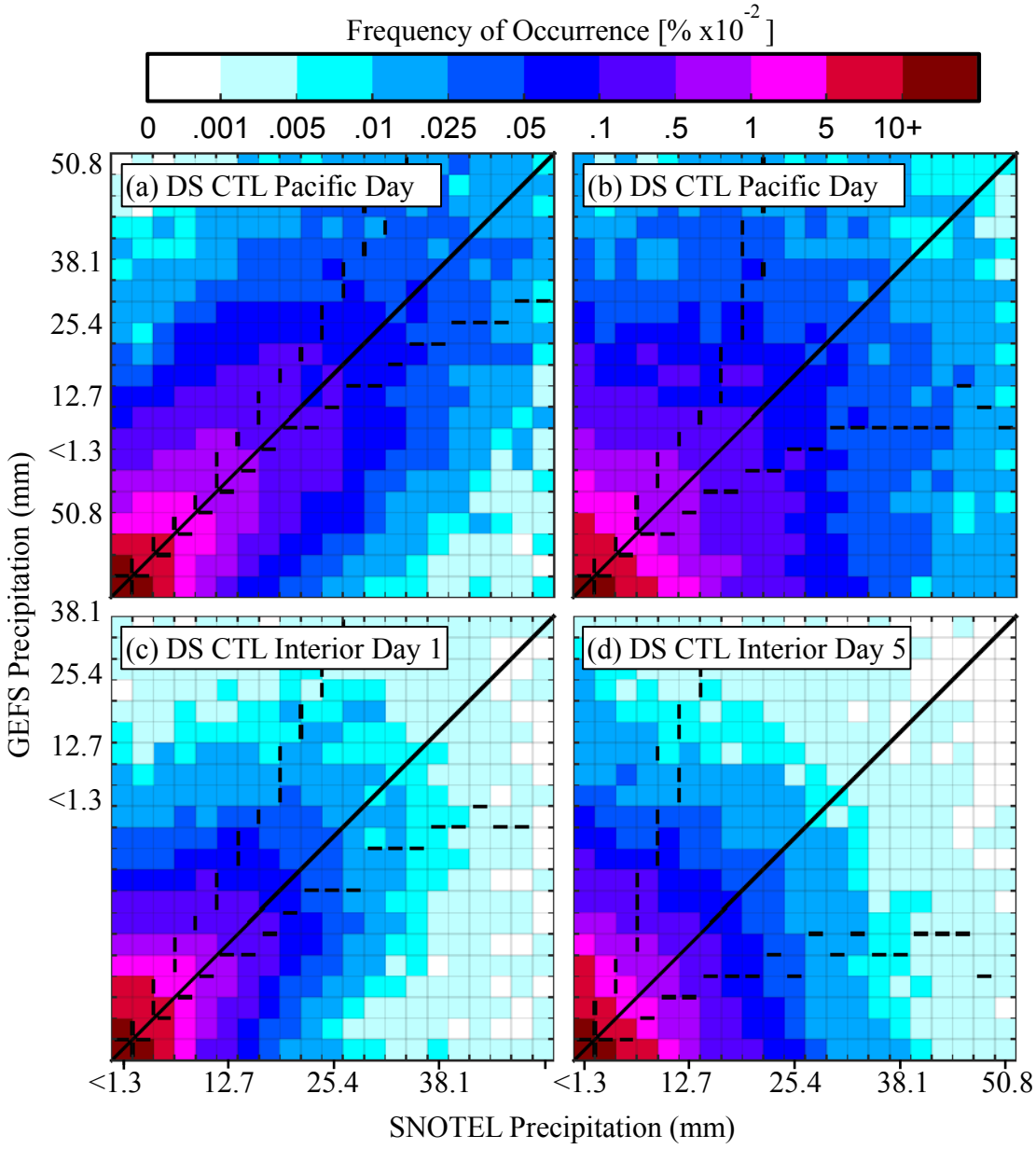

Fig. 3.9. Same as Fig. 3.6 except for the downscaled GEFS.

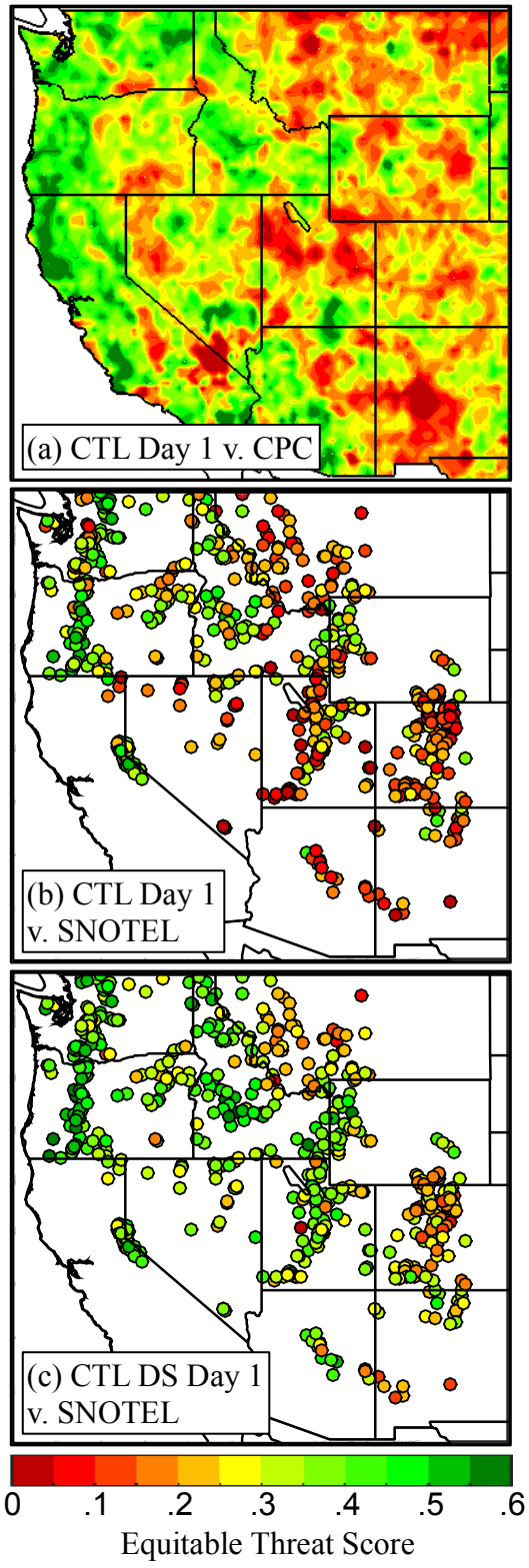

Fig. 3.10. Equitable threat scores for upper-quartile daily-precipitation events. (a) GEFS Day 1 (12-36 h) control forecasts (CTL) relative to the CPC analysis. (b) GEFS Day 1 CTL relative to SNOTEL observations. (c) GEFS Day 1 downscaled CTL relative to SNOTEL observations.

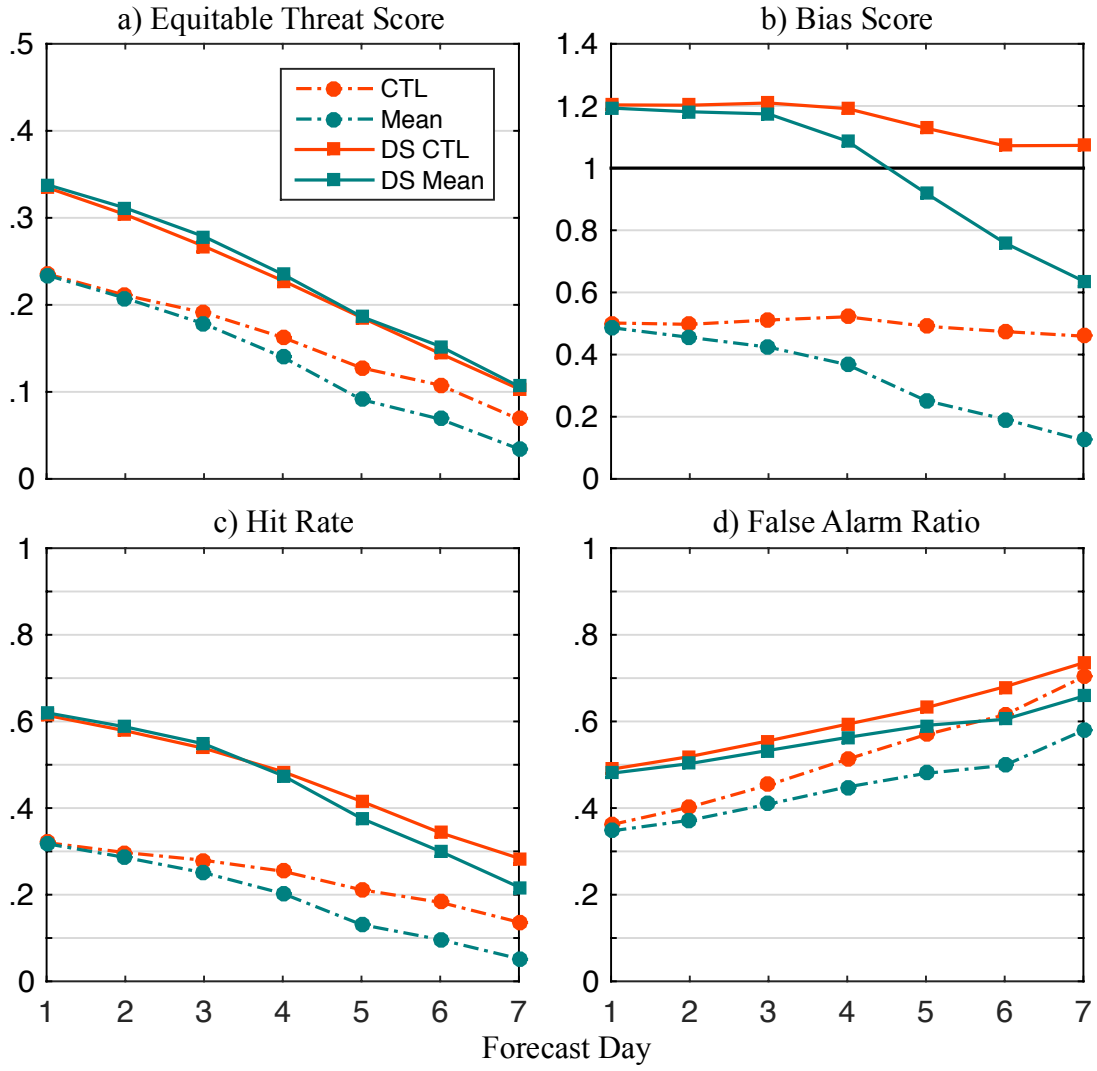

Fig. 3.11. Statistical measures of GEFS control (CTL, red dash-dot line / circle), GEFS ensemble mean (Mean, teal dash-dot line / circle), downscaled GEFS control (DS CTL, red line / square), and downscaled GEFS ensemble mean (DS Mean, teal line / square) forecasts of upper-quartile precipitation events at SNOTEL stations with increasing forecast projection. (a) Equitable threat score. (d) Bias score. (c) Hit rate. (d) False alarm ratio.

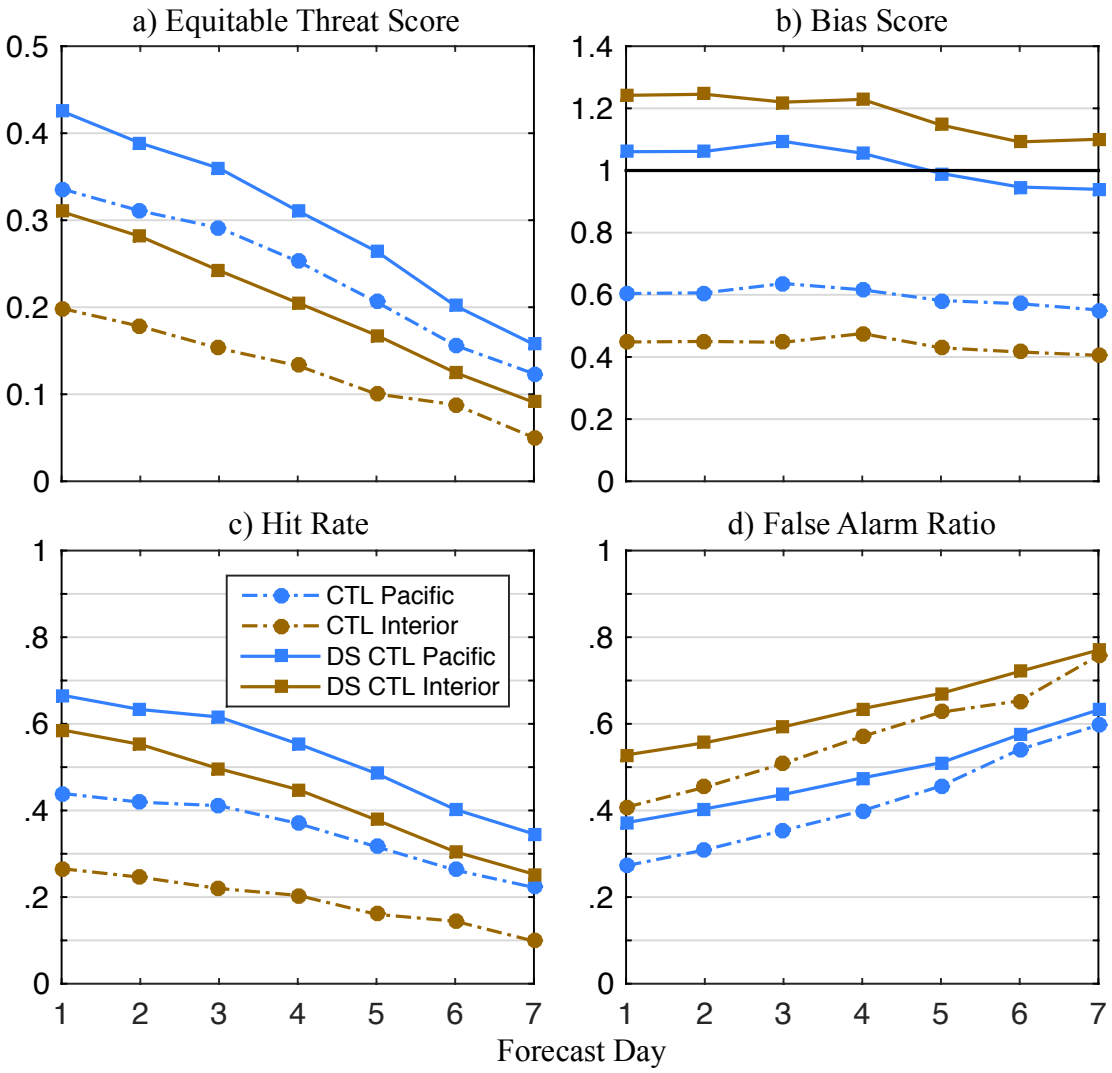

Fig. 3.12. Regional statistical measures of GEFS control (CTL, dash-dot line / circle) and downscaled GEFS control (DS CTL, line / square) forecasts for upper-quartile precipitation events at Pacific (blue) and Interior (brown) SNOTEL stations with increasing forecast projection. (a) Equitable threat score. (d) Bias score. (c) Hit rate. (d) False-alarm ratio.

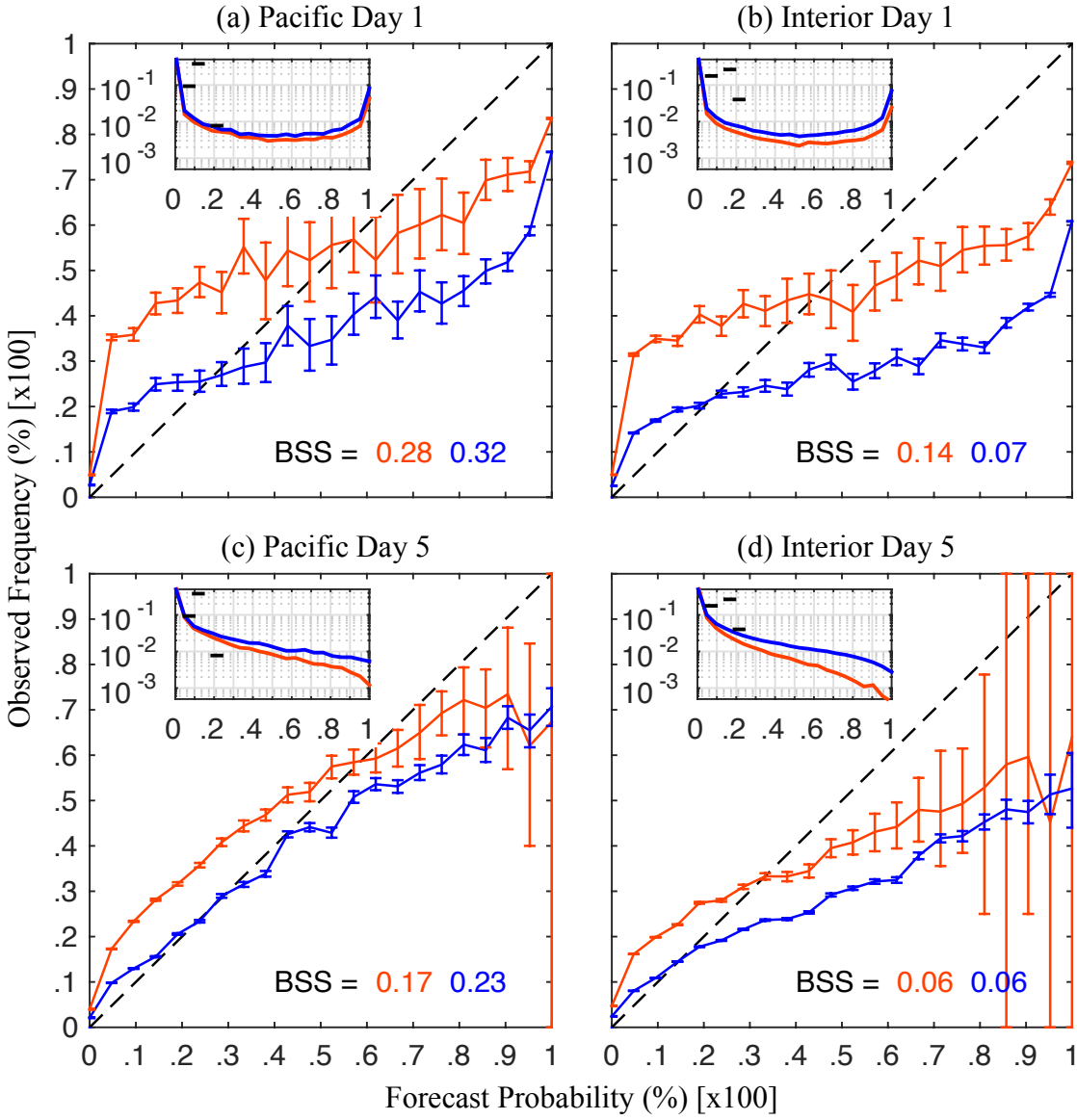

Fig. 3.13. Reliability diagrams for GEFS Day 1 (12-36 h) (red) and downscaled GEFS Day 1 (blue) forecasts of upper-quartile events at (a) Pacific and (b) Interior Ranges SNOTEL stations. (c), (d) Same as (a), (b) except for Day 5 (108-132 h). Brier skill scores (BSS) annotated. Inset histograms indicate the relative frequency of forecast probabilities for GEFS (red) and downscaled GEFS (blue) forecasts, and the frequency of climatological reference probabilities (black lines).

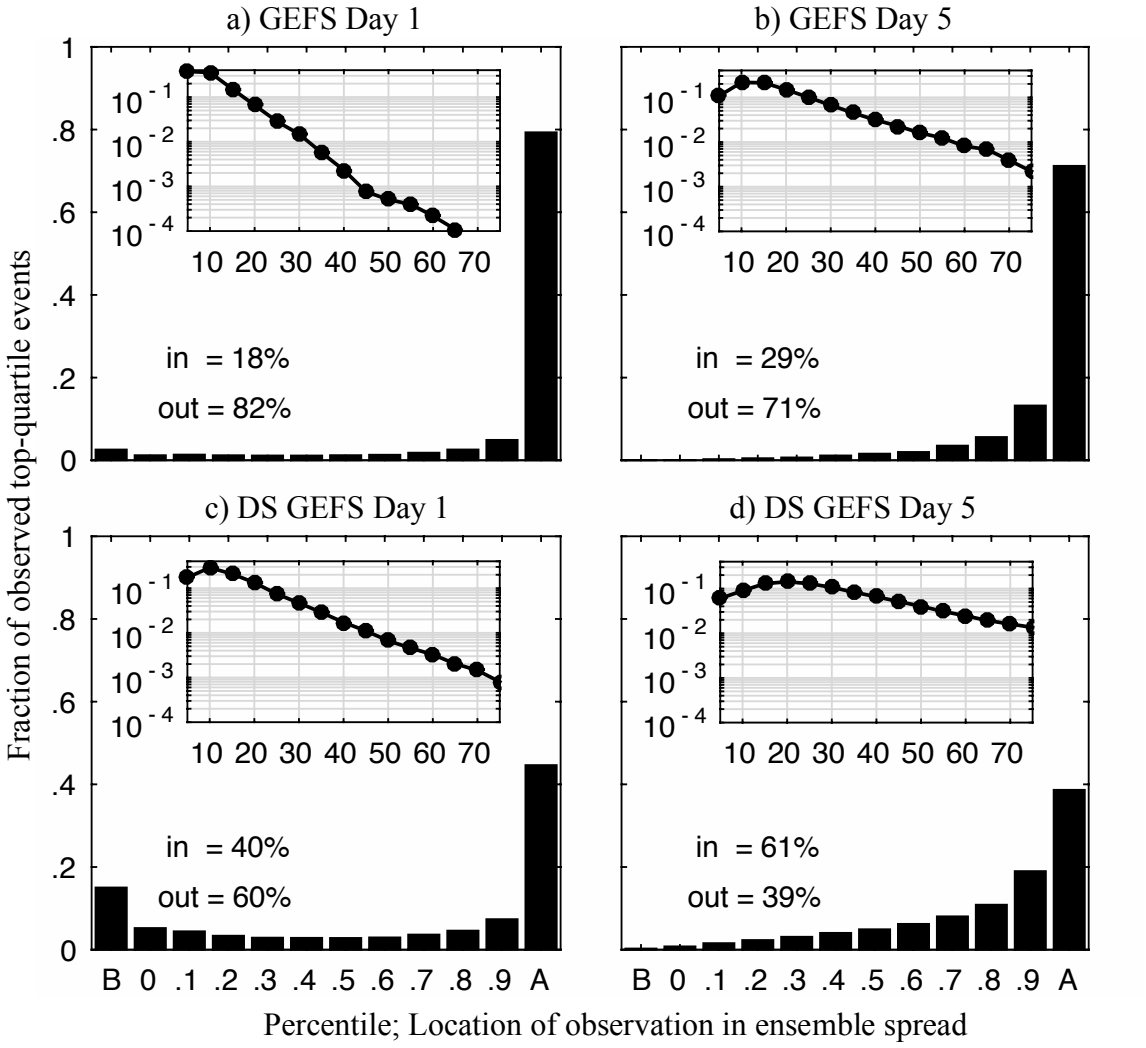

Fig. 3.14. Rank histograms for (a) GEFS Day 1 (12-36 h), (b) downscaled GEFS Day 1, (c) GEFS Day 5 (108-132 h), and (d) downscaled GEFS Day 5 forecasts of upper-quartile precipitation events observed at all SNOTEL stations. "A" and "B" indicate above and below the ensemble spread, respectively. Insert histograms indicate the frequency of ensemble spread size in 5 mm bins. Annotations of "in" and "out" reflect percent of observations occurring inside and outside of the ensemble spread, respectively.

CHAPTER 4

CONCLUSION

We have evaluated three cool-seasons (Oct–Mar) of reforecasts and forecasts produced by the operational GEFS over the western U.S. using the CPC analysis to identify broad regional biases and SNOTEL observations for gauge-based validation in upper-elevation regions. Validation against the CPC precipitation analysis shows that the GEFS control (as well as individual members) generally produces too little precipitation over and upstream of topographic barriers and too much precipitation in downstream valleys and basins. Relative to SNOTEL observations, which are preferentially located in relatively wet upper-elevation areas, the GEFS control (and other members) has a pronounced dry bias at most locations. This dry bias reflects the infrequent production of larger 24-h precipitation events [i.e., $\geq$ 22.9 mm (10.2 mm) at stations in the Pacific (Interior) Ranges] relative to observations. Bivariate histograms show that at all but the smallest thresholds, observed events are more likely than not to be underforecast, with the likelihood increasing toward the interior.

For traditional performance measures [e.g., equitable threat score (ETS), hit rate, bias score, and false alarm ratio], the performance of the GEFS control (and other members) for upper-quartile precipitation events is highest in the Pacific Ranges and generally degrades toward the interior with considerable spatial variability. Based on

ETS, a Day 5 forecast over the Pacific Ranges is as skillful as a Day 1 forecast over the Interior Ranges. Hit rates and false alarm ratios are best at Day 1, when the GEFS control upper-quartile-event hit rate is 0.44 (0.27) in the Pacific (Interior) Ranges, and the false alarm ratio is 0.27 (0.41).

Probabilistic verification statistics reflect both the underprediction biases inherent in the GEFS control (and individual members), as well as the unreliable (or overconfident) and underdispersive nature of the GEFS. Observed upper-quartile precipitation events at SNOTEL stations exceed the wettest member of the GEFS ensemble at Day 1 (Day 5) ~80% (~70%) of the time. At Day 1, PQPFs for upper-quartile events are strongly overconfident, with low-probability (high-probability) forecasts associated with a higher (lower) frequency of observed events. Reliability improves with increasing forecast lead time, but high-probability forecast overconfidence is still evident at Day 5. Forecasters should be aware that although the GEFS has a low frequency bias for larger events, a high PQPF for a larger event is likely an overestimate of the actual event probability. Day 1 and Day 5 PQPFs for upper-quartile events in the Pacific Ranges are somewhat more skillful than using climatological probabilities (BSS=0.28 and 0.17, respectively), but over the Interior Ranges, such PQPFs are about as skillful as climatological probabilities (BSS=0.14 and 0.06, respectively).

In an attempt to rescale the low-resolution GEFS forecasts and reduce the impact of GEFS biases at SNOTEL stations, we produced statistically downscaled forecasts derived from high-resolution climatological precipitation analyses produced by the PRISM Climate Group at Oregon State University. Such downscaling generally resolves dry biases at SNOTEL locations, as well as the tendency for most events to be

underforecast. These improvements largely remove event frequency biases in the Pacific Ranges, while event frequencies are greater than observed for events ≥ 10.2 mm in the Interior Ranges. Downscaling also improves ETSs, hit rates, and bias scores. For example, the downscaled GEFS control ETS for upper-quartile events at Day 4 is roughly equivalent to that of the undownscaled GEFS control at Day 1. However, upper-quartile-event false alarm ratios at Day 1 are worsened to 0.37 (0.53) in the Pacific (Interior) Ranges.

For PQPFs, downscaling worsens reliably by exacerbating the overconfidence of high-probability forecasts. However, at Day 1 (Day 5), 40% (61%) of upper-quartile events are captured by the downscaled ensemble spread, which is an improvement over the undownscaled GEFS. Nevertheless, most missed events are underforecast by the wettest ensemble member (rather than overforecast by the driest member), despite the overprediction issues in the downscaled GEFS. Downscaled PQPFs in the Pacific Ranges have slightly improved BSSs, while downscaled PQPFs in the Interior Ranges do not have improved BSSs.

These findings indicate that the GEFS lacks sufficient resolution to reliably produce QPFs and PQPFs over mountainous terrain in the western U.S. where orographic effects are prominent. Such forecasts are particularly problematic for larger events over the fine-scale topography of the western interior. Efforts to improve these forecasts through climatology-based downscaling yield some improvements, but also increase false alarms, especially over the interior. The extent to which these results are exacerbated by the relatively low-resolution 1.0° grid is unclear and perhaps some improvement would occur with a higher resolution output grid. However, even at native grid spacing (~33

km), fine-scale orographic effects remain unresolved. Western U.S. forecasters should be aware of the capabilities and limitations of the GEFS and downscaled GEFS identified herein over the western U.S. Future work should examine the performance of alternative downscaling and ensemble calibration approaches as these offer a pathway to improved forecasts as long as ensemble grid spacing fails to resolve key orographic precipitation processes.

REFERENCES

Avanzi, F., C. D. Michele, A. Ghezzi, C. Jommi, and M. Pepe, 2014: A processing-modeling routine to use SNOTEL hourly data in snowpack dynamic models. *Adv. Water Resour.*, **73**, 16-29.

Baxter, M. A., G. M. Lackmann, K. M. Mahoney, T. E. Workoff, and T. M. Hamill, 2014: Verification of quantitative precipitation reforecasts over the southeastern United States. *Wea. Forecasting*, **29**, 1199-1207.

Black, A. W., and T. L. Mote, 2015: Characteristics of winter-precipitation-related transportation fatalities in the United States. *Wea. Climate Soc.*, **7**, 133-144.

Bontron, G., and C. Obled, 2005: A probabilistic adaptation of meteorological model outputs to hydrological forecasting. *Houille Blanche-Rev. Int. Eau,* **1**, 23-28.

Brocker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651-661.

Charles, M. E., and B. A. Colle, 2009: Verification of extratropical cyclones with the NCEP operational models. Part 1: Analysis errors and short-term NAM and GFS forecasts. *Wea. Forecasting*, **24**, 1173-1190.

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. W. Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.,* **113**, D04110.

Cohen, J., 1996: Snowstorms. *Encyclopedia of Weather and Climate*, S. H. Schneider, Ed., Vol. 2, Oxford University Press, 700-703.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140-158.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031-2064.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting,* **13**, 1132-1147.

Fassnacht, S. R., 2004: Estimating Alter-shielded gauge snowfall undercatch, snowpack sublimation, and blowing snow transport at six sites in the coterminous USA. *Hydrol. Process.*, **18**, 3481-3492.

Gutmann, E. D., R. M. Rasmussen, C. Liu, K. Ikeda, D. J. Gochis, M. P. Clark, J. Dudhia, and G. Thompson, 2012: A comparison of statistical and dynamic downscaling of winter precipitation over complex terrain. *J. Clim.*, **25**, 262-281.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550-560.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905-2923.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.,* **134**, 3209-3229.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2007: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev*., **136**, 2620-2632.

Hamill, T. M., 2011: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Wea. Rev.*, **140**, 2232-2252.

Haren, R. V., R. J. Haarsma, G. J. V. Oldenborgh, and W. Hazeleger, 2015: Resolution dependence of European precipitation in a state-of-the-art atmospheric general circulation model. *J. Climate*, **28**, 5134-5149.

Hatchett, B., and M. Kaplan, 2016: Some characteristics of upside-down storms in the northern Sierra Nevada, California-Nevada, USA. *J. Oper. Meteor*. Submitted.

Higgins, R. W., W. Shi, E. Yarosh, and R. Joyce, 2000: *Improved United States precipitation quality control system and analysis*. NCEP/Climate Prediction Center Atlas 7, NCEP/CPC. [Available online at http://www.cpc.ncep.noaa.gov/products/outreach/research_papers/ncep_cpc_atlas/7/.]

Hou, D., M. Charles, Y. Luo, Z. Toth, Y. Zhu, R. Krzysztofowicz, Y. Lin, P. Xie, D. J.

Seo, M. Pena, and B. Cui, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542-2557.

Ikeda, K., R. Rasmussen, C. Liu, D. Gochis, D. Yates, F. Chen, M. Tewari, M. Barlage, J. Dudhia, K. Miller, K. Arsenault, V. Grubisic, G. Thompson, and E. Guttan, 2010: Simulation of seasonal snowfall over Colorado. *Atmos. Res.*, **97**, 462-477.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons Ltd, 240 pp.

Junker, N. W., J. E. Hoke, B. E. Sullivan, K. F. Brill, and F. J. Hughes, 1992: Seasonal geographic variations in quantitative precipitation predication by NMC's nested-grid model and medium-range forecast model. *Wea. Forecasting*, **7**, 410-429.

Kunz, M., and C. Kottmeier, 2006: Orographic enhancement of precipitation over low mountain ranges. Part II: Simulations of heavy precipitation events over southwest Germany. *J. Appl. Meteor. Climatol.*, **45**, 1041-1055.

Marsigli, C., A. Montani, F. Nerozzi, T. Paccagnella, S. Tibaldi, F. Molteni, and R. Buizza, 2001: A strategy for high-resolution ensemble prediction. Part II: Limited-area experiments in four alpine flood events. *Quart. J. Roy. Meteor. Soc.*, **127**, 2095-2115.

Mullen S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **129**, 638-663.

Neiman, P. J., L. J. Schick, F. M. Ralph, M. Hughes, and G. A. Wick, 2011: Flooding in western Washington: The connection to atmospheric rivers. *J. Hydometeor.*, **12**, 1337-1358.

NOAA, 2015: Technical Implementation Notice 15-43. Accessed 3 September 2015. [Available online at http://www.nws.noaa.gov/os/notification/tin15-43gefs.htm.]

Parker, L. E., and J. T. Abatzoglou, 2016: Spatial coherence of extreme precipitation events in the northwestern United States. *Int. J. Climatol.*, **36**, 2451-2460.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155-1174.

Ralph, F. M., P. J. Neiman, G. A. Wick, S. I. Gutman, M. D. Dettinger, D. R. Cayan, and A. B. White, 2006: Flooding on California's Russian River: Role of atmospheric rivers. *Geophys. Res. Lett.*, **33**, L13801.

Rutledge, G. K., J. Alpert, and W. Ebisuzaki, 2006: NOMADS: A climate and weather

model archive at the National Oceanic and Atmospheric Administration. *Bull. Amer. Meteor. Soc.*, **87**, 327-341.

Rutz, J. J., and W. J. Steenburgh, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142**, 905-921.

Rutz, J. J., W. J. Steenburgh, and F. M. Ralph, 2015: The inland penetration of atmospheric rivers over western North America: A Lagrangian analysis. *Mon. Wea. Rev.*, **143**, 1924-1944.

Schaefer J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570-575.

Schirmer M., and B. Jamieson, 2015: Verification of analyzed and forecasted winter precipitation in complex terrain. *The Cryosphere*, **9**, 587-601.

Serreze, M. C., M. P. Clark, R. L. Armstrong, D. A. McGinnis, and R. S. Pulwarty, 1999: Characteristics of the western United States snowpack telemetry (SNOTEL) data. *Water Resour. Res.*, **35** (7), 2145–2160.

Serreze, M. C., M. P. Clark, and A. Frei, 2001: Characteristics of larger snowfall events in the montane western United States as examined using snowpack telemetry (SNOTEL) data. *Water Resour. Res.*, **37** (5), 675-688.

Smith, B. L., S. E. Yuter, P. J. Neiman, and D. E. Kingsmill, 2010: Water vapor fluxes and orographic precipitation over northern California associated with a landfalling atmospheric river. *Mon. Wea. Rev.*, **138**, 74-100.

Steenburgh, W. J., 2003: One hundred inches in one hundred hours: Evolution of a Wasatch Mountain winter storm cycle. *Wea. Forecasting*, **18**, 1018-1036.

Steenburgh, W. J., 2004: One hundred inches in one hundred hours – the complex evolution of an intermountain winter storm cycle. *Bull. Amer. Meteor. Soc.*, **85**, 16-20.

Steenburgh, W. J., 2014: *Secrets of the Greatest Snow on Earth*. Utah State University Press, 244 pp.

Stensrud, D. J., H. E. Brooks, J. Du, S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433-446.

Stewart, R. E., D. Bachand, R. R. Dunkley, A. C. Giles, B. Lawson, L. Legal, S. T. Miller, B. P. Murphy, M. N. Parker, B. J. Paruk, and M. K. Yau, 1995: Winter storms over Canada. *Atmos.-Ocean*, **33** (2), 223-247.

Tremper, B., 2008: *Staying Alive in Avalanche Terrain*. Mountaineers Books, 318 pp.

U. S. Department of the Interior, 2012: Flood of January 1997 in the Truckee River Basin, Western Nevada. USGS, 2 pp. [Available online at http://pubs.usgs.gov/fs/1997/0123/report.pdf.]

Wilby, R. L., T. M. L. Wigley, D. Conway, P.D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks, 1998: Statistical downscaling of general circulation model output: A comparison of methods. *Water Resour. Res.*, **34** (11), 2995-3008.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd ed. Academic Press, 627 pp.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.,* **107**, 4429.

Wood, A., L. Leung, V. Sridhar, and D. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Clim. Change*, **62**, 189-216.

Xie, P., A. Yatagai, M. Chen, T. Hayasaka, Y. Fukushima, C. Liu, and S. Yang, 2007: A gauge-based analysis of daily precipitation over East Asia. J. Hydrometeor., 8, 607-626.

Yang, D., B. E. Goodison, J. R. Metcalfe, V. S. Golubev, R. Bates, T. Pangburn, and C. L. Hanson, 1998: Accuracy of NWS 8" standard nonrecording precipitation gauge: Results and application of WMO intercomparison. *J. Atmos. Oceanic Technol*., **15**, 54-68.

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Weath. Rev.*, **133**, 279-294.